PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association

Atlanta, GA; 3-5 November 2016

Version: 2 November 2016

PhilSci ARCHIVE

Table of Contents

CENTRALISED FUNDING AND THE DIVISION OF COGNITIVE LABOUR

Abstract. Project selection by funding bodies directly influences the division of cognitive labour in scientific communities. I present a novel adaptation of an existing agent-based model of scientific research, in which a central funding body selects from proposed projects located on an epistemic landscape. I simulate four different selection strategies: selection based on a god's-eye perspective of project significance, selection based on past success, selection based on past funding, and random selection. Results show the size of the landscape matters: on small landscapes historical information leads to slightly better results than random selection, but on large landscapes random selection greatly outperforms historically-informed selection.

Word count: 4359

Introduction

National funding bodies support much of contemporary science. The selection criteria for funding have gained increasing attention within philosophy of science (Gillies, 2008; O'Malley et al., 2009; Haufe, 2013; Lee, 2015). Meanwhile, there has been growing interest in model-based approaches to understanding the social epistemic activities of scientists (Kitcher, 1990; Strevens, 2003; Weisberg and Muldoon, 2009; Grim, 2009; Zollman, 2010). The current paper builds on previous modelling tools to explore the effects of centralised selection mechanisms on the division of cognitive labour and the ability of scientific communities to efficiently discover significant truths.

Science aims at discovering significant truths, i.e. not just any truths, but truths that will eventually contribute in a meaningful way to well-being (Kitcher, 2001). This is the justification for the public support of science, including basic science (Bush, 1945). Some funding terminology: scientific projects have high *impact* (ex post) if they result in significant truths; projects have high *merit* (ex ante) if they are predicted to have high impact.

Polanyi (1962) analysed merit as being composed of three components: scientific value, plausibility and originality. Polanyi notes an essential tension between plausibility and originality: the more original a project, the more difficult it is to evaluate its plausibility. Polanyi advocates selection by peer review as a conformist position, that sacrifices the occasional meritorious original project while ensuring all supported research projects are plausible, to "prevent the adulteration of science

by cranks and dabblers" (p. 8). Gillies (2008, 2014) takes an opposing position, arguing that the cost of losing (infrequent) highly original and meritorious research is much greater than the cost of occasionally supporting implausible research that ends up being of low impact. As an alternative to peer review, Gillies advocates random selection. The tension between plausibility and originality is clearly relevant to questions of effective division of cognitive labour, and has direct links to science policy. This tension, and its complexity, is explored in this paper.

I will argue that the results of the simulations presented are both significant and surprising. The simulations show that, under reasonable parameter values for at least some fields of science, choosing projects at random performs significantly better, in terms of accumulated significant truths, compared to other funding strategies, including project selection by peer review. The results support, to an extent, Gillies' proposal of funding by lottery.

1. Model description

The model explores the influence of different funding mechanisms on the accumulation of significant truths. It builds on the epistemic landscape model developed by Weisberg and Muldoon (2009), extending it by adding representations of centralised funding selection and dynamic changes in project merit. The latter is added to reflect a more realistic picture of scientific merit. For example, Strevens (2003) discusses the effect of a successful discovery on all further pursuits of the same question: they no longer have any merit, as they lose all originality. Several dynamic processes affecting merit are detailed later in the paper.

The model represents a population of scientists exploring a topic of scientific interest. They are all funded by the same central funding body to pursue projects of varying duration, measured in years. Each project's significance is allocated in advance by the modeller, from a "god's-eye" perspective. When grants end scientists successfully complete their project. Their projects' results contribute to the collection of significant truths in the field's corpus of knowledge. Funding mechanisms are compared by their ability to generate this accumulation of significant truths.

For simplicity, scientists in the model (unrealistically) do not share their findings nor explore similar projects during research. They only work on the project for which they were funded and they only share their results at the end of a grant. The social processes set aside here have been explored in previous works (Grim, 2009; Zollman, 2010). Future work may combine the different models towards a unified picture of the division of cognitive labour.

Funding is represented as a process of selection. In every time step, the scientists whose grants have run out are placed in a pool of candidates along with new entrants to the field, and the modelled funding mechanism selects from this pool of candidates those who will receive funding and carry out research projects. Modelled funding mechanisms differ in the way they select individuals, as outlined below.

Actual potential: Actual potential, which can only be known from a god's-eye perspective, is the significance of a project's results *were it successfully completed today*. In the absence of time-dependant merit, actual potential is simply the significance of the project's results. However, in the presence of time-dependence the significance could change between the initiation of the project (at the point of funding) and its completion (at the point of contributing the results to the relevant corpus). This means that in the presence of time-dependence, actual potential might diverge from the eventual contribution of the project.

Estimated potential: Estimated potential is the scientific community's ex ante evaluation (assumed, for simplicity, to be single-valued) of the merit of a proposed project. This prediction is taken to rely on the known contributions of past projects which bear some similarity to the proposed project, and so depends on the history of research projects in the field. In representing decisions based on the research community's prediction, this selection method is akin to peer-review.

Past funding: Under this mechanism, funding is allocated to those scientists who already received funding in the past, and only to them. The model (unrealistically) represents all scientists as being of equal skill, and so this mechanism cannot be taken to mean the selection of the most "intrinsically able" scientists. Rather, this mechanism is included as a "most conservative" option, not admitting any new researchers to the field beyond the field's original investigators.

Lottery: Under a lottery, all candidates have equal chances of being funded. The lottery option serves both as a natural benchmark for other funding methods, and as a representation of the mechanism proposed by Gillies (2014).

The essence of the model is the comparison of the performance of these selection mechanisms in generating results of high significance over time under various conditions.

To represent in the model the time-dependence of merit, the significance contributions of different project results are allowed to change over time as a response to scientists' actions. Three dynamic processes are included in the model (details in §2.5). Two processes involve a reduction of significance following a successful project or breakthrough,

which reflects the one-off nature of discovery. The third process involves an increase in significance when a new avenue of research is opened by a significant discovery. Simulations based on the model show that these dynamic processes have a significant effect on the relative performance of different funding strategies.

2. Simulation details

2.1. **Simulating the epistemic landscape.** To investigate the complex nature of the domain being modelled, the model was turned into a computer simulation.[1] The basic structure of the landscape simulation follows Weisberg and Muldoon's, of a two-dimensional configuration space, charted with two coordinates x and y, with an associated scalar field represented in a third dimension as height along the z axis. Each (x, y) coordinate pair specifies a different potential research project; the closer two projects are on the landscape, the more similar they are. The scalar value associated to the coordinate represents the significance of the result obtained on a successful completion of the project, were it completed today (allowing for time dependence). The limit to two spatial dimensions of variation between projects is likely to be unrealistic (Wilkins, 2008), but a higher-dimensional alternative would make the model much less tractable.

In each run of the simulation, the landscape is generated anew in the following process:

(1) Initialise a flat surface of the required dimensions.
(2) Choose a random location on the surface.
(3) Pick random values for relative height, width along x, and width along y.
(4) Add to the landscape a hill at the location chosen in step 2 by using a bivariate Gaussian distribution with the parameters picked in step 3.
(5) Repeat steps 2-4 until the specified number of hills is reached.
(6) Scale up linearly the height of the landscape according to the specified maximum height.

This process generates the "god's-eye" perspective of the research potential of the domain. Here and later, random variables are used to fill-in parameters whose existence is essential for the simulation, but where (1) the specific values they take can vary across a range of valid model targets, and/or (2) there is no compelling empirical evidence to choose a particular value. This requires, however, several runs of the simulation for each configuration, to average out the effects of random variation.

[1] Source code for the simulation is available from the author on request.

2.2. **Simulating agents.** The agents in the model represent scientists investigating the epistemic landscape. Each agent represents an independent researcher or group, and is characterised by its location on the landscape, representing the project they are currently pursuing, and a countdown counter, representing the time remaining until their current project is finished. Like Weisberg and Muldoon's "hill climbers", agents are simulated as local maximisers. Agents follow the following strategy every simulation step:

(1) Reduce countdown by 1.
(2) If countdown is not zero: remain in same location.
(3) If countdown is zero: contribute to the accumulated significance the significance of the current location, and attempt to move to the highest local neighbour.

In the simulation, the agents are identical, in the sense that any agent, when successfully completing a project of a given significance, will contribute exactly that amount to the accumulated significance of the field. This simplification ignores natural ability and gained experience, and stems from a focus on a particular approach to science funding, which funds *projects*, rather than funding *people*. The focus is informed by the explicit policies of certain funding bodies, like the National Institutes of Health (NIH), reflected, for example, in the institution of blind peer review. Thus, the results of the current work would not extend to the minority of science funding bodies, such as the Wellcome Trust, that make explicit their preference to fund people rather than projects.

The *local neighbourhood* of an agent is defined as the 3×3 square centred on their current position. The attempt to move to the highest neighbour depends on the selection (funding) mechanism, as discussed below. The *accumulated significance*, which is the sum of all individual contributions to significance, is stored as a global variable of the simulation and used to compare strategies.

In the beginning of the simulation, a specified number of agents are seeded in random locations on the landscape, with randomly generated countdowns selected from a specified range of values. An example of an initial seeding of agents can be seen in Fig. 1.

In the absence of selection and time-dependence, the course of the simulation is easy to describe: agents begin in random locations on a random landscape, and as the simulation progresses the agents finish projects and climb local hills, until, after an amount of time which depends on the size of the landscape, the number and size of peaks, and the duration of grants, all agents trace a path to their local maxima and stay there. Since agents increase their local significance during the climb, the rate of significance accumulation increases initially, until all agents reach their local maxima, at which point significance continues accumulating at a fixed rate indefinitely. This is the dynamic

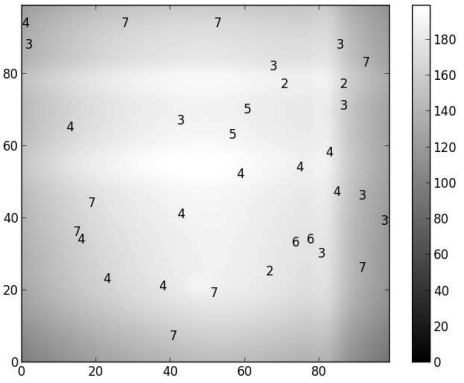


FIGURE 1. Landscape simulation with initial seeding of agents. Each number on the landscape represents an agent at its location, with the value of the number representing the agent's countdown. The colours indicate the height (significance) of each position (project) in the landscape.

seen in Weisberg and Muldoon's simulation for a pure community of "hill climbers", and its unrealistic nature highlights the importance of simulating the time-dependence of significance.

2.3. **Simulating communal knowledge.** In addition to their contribution to significance, agents also contribute to the *visibility* of the landscape (Muldoon and Weisberg, 2011). The visibility of a project represents whether the scientific community, and especially funding bodies, can estimate the significance contribution of that project. Initially, the entire landscape is invisible, representing full uncertainty. Upon initial seeding of agents, each agent contributes vision of their local neighbourhood, as defined above, to the total vision. As the agents move, they add vision of their new local neighbourhood. Visibility is used in the *best_visible* funding mechanism described below.

The simulation represents visibility in a simplistic manner by assigning binary values: either the community knows what the significance of a project will be, or it does not. A more realistic representation will allow partial visibility, with some distance decay effect, such that the community would still be able to make predictions of significance for less familiar projects, but these predictions will have a probability of being wrong, with the probability of error increasing the more unfamiliar these projects are. This addition, however, will be computationally heavy, as it requires maintaining multiple versions of the landscape, both for the real values and for the estimated values.

2.4. **Simulating funding strategies.** The aim of the model is to explore the effects of funding mechanisms on the population and distribution of investigators. Since the aim is to simulate current funding practices (albeit in a highly idealised manner), and since current funding practices operate in passive mode (choosing from proposals originating from scientists rather than dictating which projects aught be pursued), the guiding principle of the simulation is that a funding mechanism is akin to a selection process: at each step of the simulation, the actual population of agents is a subset of the candidate or potential population, where inclusion in the actual population follows a certain selection mechanism.

Funding mechanisms are simulated in the following manner:
Every step:

(1) Place all agents with zero countdown in a pool of "old candidates".
(2) Generate a set of new candidate agents, in a process identical to the seeding of agents in the beginning of the simulation.
(3) Select from the joint pool of (old candidates + new candidates) a subset according to the selection mechanism specified by the funding method.
(4) Only selected agents are placed on the landscape and take part in the remainder of the simulation, the rest are ignored.

The simulation can represent four different funding mechanisms:

best: selects the candidates which are located at the highest points, regardless of the visibility of their locations. This simulates a mechanism which selects the most promising projects from a god's eye perspective. This overly optimistic mechanism does not represent a real funding strategy. Rather, it serves as an ideal benchmark against which realistic funding mechanisms are measured.

best_visible: filters out candidates which are located at invisible locations, i.e. candidates who propose to work on projects which are too different from present or past projects. It then selects the candidates in the highest locations from the remainder. This strategy is closer to a realistic representation of selection by peer review. Note that even this version is epistemically optimistic, as it assumes the selection panel has successfully gathered all available information from all the different agents, both past and present.

lotto: selects candidates at random from the candidate pool, disregarding the visibility and height of their locations.

oldboys: represents no selection: old candidates continue, no new candidates are generated.

The key parameters for all funding mechanisms are the size of the candidate pool and the size of the selection pool. The size of the candidate pool, which in turn depends on the size of the new candidate pool (as the size of the old candidate pool emerges from the simulation), has been chosen in the simulations such that the total candidate pool is equal in size to the initial number of agents (except *oldboys* where there are no new candidates). This means the success probability changes between funding rounds, around a mean which is equal to 1/(average countdown). With an average grant duration of five years, this yields a success rate of 20%, close to the real value in many contemporary funding schemes (NIH, 2014). The number of grants awarded each year is set to equal the number of grants completed each year, maintaining a fixed size for the population of investigators.

For simplicity, the simulated funding mechanisms do not take into account the positions of existing agents on the landscape, except indirectly when considering their vision. Future simulations may consider a selection mechanism which explicitly favours either diversity or agglomeration, though one expects difficulties in operationalisation and measurement of epistemic diversity.

2.5. **Simulating merit dynamics.** To make the simulation more realistic, the significance of projects is allowed to change over time in response to research activities of the community of investigators. Three such dynamic processes are included in the simulation:

Winner takes it all: As was made explicit by Strevens (2003), the utility gain of discovery is a one-off event: the first (recognised) discovery of X may greatly contribute to the collective utility, but there is little or no contribution from further discoveries of X. In the simulation, this is represented by setting the significance of a location to zero whenever an agent at that location has finished their project and made their contribution to accumulated significance. This effect is triggered whenever any countdown reaches zero, which makes it quite common, but it has a very localised effect, only affecting the significance of a single project.

Reduced novelty: When a researcher makes a significant discovery, simulated by finishing a project with associated significance above a certain threshold, the novelty of nearby projects is reduced, which in the model is simulated by a reduction of significance in a local area around the discovery.

New avenues: When a researcher makes a significant discovery, it opens up the possibility of new avenues of research, simulated in the model by the appearance of a new randomly-shaped hill at a random location on the landscape.

3. Results and discussion

Here I present the results of simulations of different setups of interest, exploring the relative success of different funding mechanisms under different conditions.

All simulation results show a comparison between the four funding mechanisms, as a plot of total accumulated significance (arbitrary units) at the end of the simulation run, averaged over five runs with different random seeds. In all simulations the range of countdowns was 2 to 7. The number of individuals was set to equal (size of landscape)$^{3/4}$. Simulations were ran for 50 steps. The trigger for significance-dependant processes was 0.7 of the global maximum. Results are shown for a small landscape (50×50) in Fig. 2 and for a large landscape (500×500) in Fig. 3.

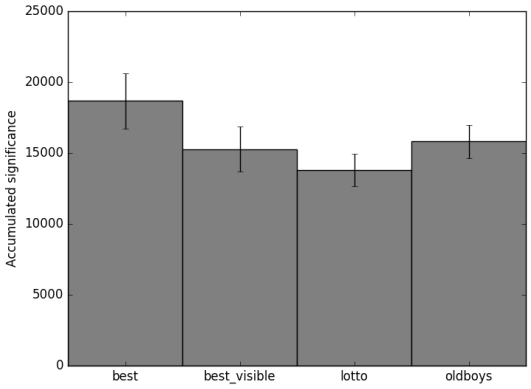


Figure 2. Comparison of significance accumulation under different funding mechanisms, small landscape (50×50).

To get a feeling for how the community is affected by the funding mechanism, I present visualisations of the state of the landscape at the end of the simulation run for the two funding mechanisms mentioned in the introduction (*best_visible* and *lotto*) in Fig. 4. Note that due to the *winner takes it all* dynamic process it is possible to "see" the past trajectory of exploration, as completed projects leave behind highly localised points of zero (remaining) significance. This allows for a visual representation of the division of cognitive labour that emerges under different funding schemes.

As is clear from the simulations, the *best* funding mechanism is indeed best at accumulating significance over time, though with various lead margins over the second best strategy. In the presence of dynamic

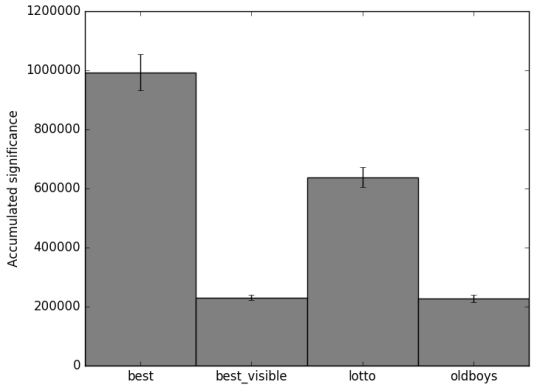
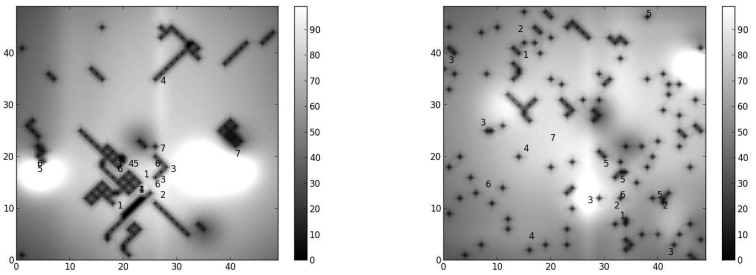


FIGURE 3. Comparison of significance accumulation under different funding mechanisms, large landscape (500×500).



(A) *best_visible* (B) *lotto*

FIGURE 4. Landscape visualisation at the end of the simulation run under different funding mechanisms.

processes, *best* is in the best position to locate new avenues for research, wherever they show up. However, as mentioned above, the *best* funding strategy is not realisable, as it requires a god's eye view of the epistemic landscape.

On the small landscape the three strategies, *best_visible*, *oldboys*, and *lotto* perform roughly similarly, with *lotto* at a small disadvantage as it cannot make use of valuable information from past successes. It seems counter-intuitive that *best_visible* performs worse than *oldboys*. A possible explanation is the effect of reduced novelty: *best_visible* tends to cluster scientists around the most promising projects, and so when one makes a breakthrough it reduces the significance of contributions for all groups working on similar projects (the phenomenon known in

contemporary science as "scooping"). This excessive clustering around fashions is not present in *oldboys* or *lotto*.

On the large landscape *lotto* greatly outperforms *best_visible* and *oldboys*. This is because new avenues on a large landscape are likely to spawn outside the visibility of the agents, where *lotto* can access them but the other two strategies cannot. In the smaller landscape this effect is not apparent, as the relative visibility is larger, and therefore the chance of a new avenue appearing within the visible area is larger.

Conclusion

This paper presented a way to extend existing epistemic landscape models so that they can represent selection by a central funding body and time dependence of significance. This model was used in computer simulations to compare the effectiveness of different idealised versions of selection criteria, most notably selection based on past successes (akin to peer review), random selection and no selection. The most significant result from the simulation was that on a large landscape, when a topic can be explored in many ways that could be very different from each other, random selection performs much better than selection based on past performance.

This result fits in with a general result from the body of works on agent-based models of scientific communities, that shows diversity in the community trumps individual pursuit of excellence as a way of making communal epistemic progress. The tension of science funding, between originality and plausibility, is thus a part of the broader tension between diversity and excellence, between exploration and exploitation.

Previous social epistemology models have focused on the role of *internal* factors in shifting the balance between exploration and exploitation. Kitcher (1990); Strevens (2003) look at reward structures (of internal credit, not external monetary rewards) and individual motivation towards credit or truth. Grim (2009); Zollman (2010) look at information availability and information transfer between scientists, and at individual beliefs. Weisberg and Muldoon (2009) look at individual researchers' social strategy: follower or maverick.

The current work is the first within this modelling lineage to look at the effects of an *external*, *institutional* factor: selection by a centralised funding body. The current paper brings this line of research closer to having a direct relevance to science policy. Hopefully future work in this vain will continue this trend, to deliver on the challenge set out by Kitcher (1990, p. 22):

> How do we best design social institutions for the advancement of learning? The philosophers have ignored the social structure of science. The point, however, is to change it.

We could start by advocating for funding mechanisms that allow for more exploration.

References

Bush, V. (1945). *Science, the endless frontier: A report to the President*. Washington: U.S. Government printing office.

Gillies, D. (2008). *How should research be organised?* London: College Publications.

Gillies, D. (2014). Selecting applications for funding: why random choice is better than peer review. *RT. A Journal on research policy and evaluation 2*(1).

Grim, P. (2009). Threshold phenomena in epistemic networks. In Complex adaptive systems and the threshold effect: Views from the natural and social sciences: Papers from the AAAI Fall Symposium, pp. 53–60.

Haufe, C. (2013). Why do funding agencies favor hypothesis testing? *Studies in History and Philosophy of Science Part A 44*(3), 363–374.

Kitcher, P. (1990). The division of cognitive labor. *The journal of philosophy 87*(1), pp. 5–22.

Kitcher, P. (2001). *Science, truth, and democracy*. New York: Oxford University Press.

Lee, C. J. (2015). Commensuration bias in peer review. *Philosophy of Science 82*(5), 1272–1283.

Muldoon, R. and M. Weisberg (2011). Robustness and idealization in models of cognitive labor. *Synthese 183*(2), 161–174.

NIH (2014). Success rates - NIH research portfolio online reporting tools (RePORT). `http://report.nih.gov/success_rates/`, Accessed 11 July 2014.

O'Malley, M. A., K. C. Elliott, C. Haufe, and R. M. Burian (2009). Philosophies of funding. *Cell 138*(4), 611–615.

Polanyi, M. (1962). The republic of science: Its political and economic theory. *Minerva 1*, 54–73.

Strevens, M. (2003). The role of the priority rule in science. *The journal of philosophy 100*(2), 55–79.

Weisberg, M. and R. Muldoon (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of science 76*(2), 225–252.

Wilkins, J. S. (2008). The adaptive landscape of science. *Biology and philosophy 23*(5), 659–671.

Zollman, K. J. (2010). The epistemic benefit of transient diversity. *Erkenntnis 72*(1), 17–35.

How To Make Selective Realism More Selective (and More Realist Too)

Massimiliano Badino
Massachusetts Institute of Technology — Universitat Autonoma de Barcelona

Abstract

Selective realism is the thesis that some wisely chosen theoretical posits are essential to science and can therefore be considered as true or approximately true. How to choose them wisely, however, is a matter of fierce contention. Generally speaking, we should favor posits that are effectively deployed in successful prediction. In this paper I propose a refinement of the notion of deployment and I argue that selective realism can be extended to include the analysis of how theoretical posits are actually deployed in symbolic practices.

1. Introduction

Among the several forms of realism, the so-called selective realism (SelRealism) is arguably the one that engages history of science more seriously. The driving idea of SelRealism is that, although theories as wholes are false and doomed to be abandoned, it is possible to select a certain number of theoretical posits (TPs) that are likely to be maintained in future theories and are therefore true or approximately true. How to determine these TPs is *partly* an empirical question—and this explains the historical character of the SelRealism program—but it cannot be *merely* an empirical question lest one end up in post-hoc rationalizations. A central issue of

SelRealism, hence, is how to specify criteria to properly conceptualize the TPs on which one should place one's realist commitment.

In this paper, I argue that contemporary approaches to SelRealism have neglected an important element related to the way in which theoretical claims are deployed in scientific theories (Section 2). In Section 3, I propose a refinement of SelRealism based on the distinction between deploying a TP fundamentally and deploying it in a non-accidental fashion. I use the concept of symbolic practices to articulate this distinction. Finally, in Section 4, I clarify my points by discussing the early development of perturbation theory.

2. Selective Realism: Theory and Practice

The upholders of SelRealism cherish two fundamental ambitions. First and foremost, they aim at making a good use of the so-called no-miracles argument (NMA) according to which one can justifiably infer the truth (or the approximate truth) of a successful theory, because, otherwise, the success would remained inexplicable. The NMA is considered to be the strongest support to realisms of any sort (Musgrave 1988; Psillos 1999, 68-94). A challenging objection to the NMA is the pessimistic meta-induction (PMI) originally formulated by Larry Laudan. According to this argument, the success of a theory is never a sufficient reason to infer even its approximate truth because history of science is replete with examples of very successful theories that wound up overthrown at some later stage. As it is likely the case that our most successful theories will suffer the same fate in the future, one has

to conclude that the realist commitment is not justified (Laudan 1981). Among the several responses to the PMI, one consists in noticing that the failures of past theories, in fact, did not depend on those TPs that lead them to success. In other words, granted Laudan's point that successful past theories are false as wholes, it can still be argued that the constituents of those theories that were responsible for their empirical success have been retained in our current science. Thus, the realist needs only to shift her commitment from theories as wholes to those enduring TPs that, being essential for success, can be justifiably believed to be true or approximately true.

The next question is, of course, how to determine those TPs. Thus, the second ambition of the upholders of SelRealism is to solve the problem of selectivity in some principled way and so beat the PMI. In one of the first instantiations of SelRealism, Philip Kitcher argued that one must "distinguish between those parts of theory that are genuinely used in the success and those that are idle wheels" (Kitcher 1993, 143). The point of this distinction is that credit for the success of a theory should be due only to those TPs that effectively contribute to it. Elaborating on Kitcher's intuition, one can argue that the program of SelRealism is based on two major conditions:

(S) Success condition: the selection of the important TPs must hinge on their relation with some significant success of the theory.

(D) Deployment condition: one must select those TPs that were effectively used in scoring that success.

Let me briefly comment on these two conditions. While (S) is now a realist trademark, the deployment condition (D) is what sets apart SelRealism from other forms of realism, such as structural realism, also engaged in picking out enduring elements of scientific theories (Worrall 1989; Chakravartty 2011). It is also important to notice that (S) and (D) are independent conditions. Firstly, (S) refers to a relation between the selected TP and empirical success, while (D) refers to a relation between the TP and the rest of the theory. Secondly, either condition can be satisfied separately. (D) has been added precisely to avoid those cases in which idle TPs are involved in empirical success and, obviously, there are scores of examples of TPs used by theories which however never led to any success. It follows that, while (S) is supposed to meet the first ambition of SelRealism, the second ambition, to block the PMI, is on (D).

So much for SelRealism in theory. Let us now examine how this program has been carried out in practice. One of the first philosophers to seriously elaborate on Kitcher's suggestion was Stathis Psillos. His criterion for selecting TPs works in the following way (Psillos 1999, 110). Let us assume that a certain successful prediction *P* can be obtained by combining the TPs *H*, *H*' and the auxiliaries *A*.[1] According to

[1] For virtually all writers, empirical success means "successful prediction". David Harker has leveled important criticisms against this tendency to interpret success in terms of individual predictions and has suggested that success should be understood as progress, i.e. in terms of the improvements a theory makes with respect to its predecessors (Harker 2008, 2013).

Psillos, the TP *H* is essential to success *P* and should be considered true or approximately true if and only if:

(1) *H*' and *A* alone do not lead to *P*.

(2) There is no alternative *H** to *H* such that:

(a) *H** is consistent with *H*' and *A*;

(b) *H**, *H*', and *A* lead to P;

(c) *H** is not *ad hoc* or otherwise purposefully concocted to lead to *P*.

This criterion is the bedrock of Psillos's *divide et impera* strategy. The driving intuition behind it is to capture the *indispensability* of *H*: we should place our realist commitment upon those TPs without which empirical success cannot be obtained. However, Tim Lyons has cogently argued that Psillos's criterion fails to characterize indispensability (Lyons 2006). The indispensability of *H* should be ensured by condition (2), which states, in brief, that *H* cannot be replaced by any other TP. But, Lyons notices, "there will always be other hypotheses, albeit some that we find very unappealing, from which any given prediction can be derived" (Lyons 2006, 540). More importantly, Lyons argues, Psillos's criterion is not even an effective means for credit attribution, because it does not tell us much about how *H* contributes to the empirical success *P*. In particular, condition (2) has no relevance whatsoever for *H*'s specific contribution, because it only concerns conceivable alternatives to *H*, alternatives that, if *H* is at hand, nobody would even bother to explore. Lyons

perceptively stresses that the problem with Psillos's criterion boils down to the fact that it obliterates condition (D): "by introducing his criterion, [Psillos] has discarded the central idea of deployment realism—introduced by Kitcher and seemingly advocated by Psillos himself" (Lyons 2006, 541). It is interesting to note that, by dropping condition (D), Psillos's position becomes vulnerable to another form of PMI. One could think of getting around of Lyons's first objection by arguing that, even though an alternative to *H* is always conceivable, *at the present state* of our knowledge it is not, therefore the objection is empty. In other words, one could inject the time factor in Psillos's criterion and make it a statement of our actual best knowledge. But then the PMI crops up again, because history shows that there is no guarantee that what is indispensable today will be so tomorrow. The whole point of the PMI is that there is nothing special in our knowledge as far as it is considered *present*, because there have been a lot of *present knowledges* that have been blissfully abandoned. This is why one needs condition (D): what makes our present knowledge so special is not its happening at a certain time, but its having gone through a certain *process*, i.e., a form of deployment. The fact that our present knowledge has been deployed at lengths and it is still with us constitutes a reason to believe that it is true or approximately true.

3. Deconstructing Deployment

Having grasped that the flaw in Psillos's criterion is the dropping of the deployment condition, Lyons suggests to run to the other end of the spectrum and to inflate

dramatically the notion of deployment. His "responsibility model" consists in discarding selectivity altogether and in considering responsible for the empirical success of a theory each and every element that was originally deployed: "credit will have to be attributed to all responsible constituents, including mere heuristics (such as mystical beliefs), weak analogies, mistaken calculations, logically invalid reasoning etc." (Lyons 2006, 543). Clearly, Lyons's proposal amounts to a crack-up of the entire SelRealism program. But, more importantly, I do not think that the responsibility model captures the correct significance of (D). As my previous considerations about the PMI show, the deployment condition is not merely supposed to tell us that a TP has been effectively used in obtaining empirical success (as opposed to be *dispensable*), but also that it has been robustly so (as opposed to be merely *accidental*). What makes it plausible that a TP will still play a role in future theories is the fact that its importance for empirical success has been tested by extensive and repeated deployment. It is therefore clear that there are two ideas nested in the deployment condition. One is the idea, captured by Psillos's criterion, that significant TPs must play a fundamental role in success in order to distinguish them from idle hypotheses; the other is the idea that the deployment of a TP must ensure that its success is not accidental. These are two distinct ideas. It might happen, for example, that a TP plays an essential role in deriving a prediction in virtue of fortuitous factors cancellation or other favorable circumstances. So, while an *intensive deployment* ensure the *fundamentality* of a TP, an *extensive deployment* founds its *robustness*. Both fundamentality and robustness are ways to articulate the complex relation between a

TP and the rest of the theory, or at least some parts of the theory (more on this in a bit). Further, while fundamentality is an atemporal articulation of this relation,[2] robustness concerns precisely the temporal dimension of the deployment condition that escaped Lyons's analysis: robustness, as we shall see below, is achieved over time.

In order to clarify the distinction between fundamentality and robustness, I introduce the notion of *symbolic practices*. By symbolic practices I mean all the methods customarily used in science to manipulate symbols.[3] These include, but are not limited to, mathematical methods, formal tools, approximations procedures, models, heuristics, solution tricks, and any sort of way by which one can transform a symbolic expression into another symbolic expression. Symbolic practices are the set of methods adopted by a theory to "put to work" a certain TP or, in other words, to deploy it in order to set problems and to interpret solutions. By using the concept of symbolic practices, one can reformulate the two ideas of the deployment condition in the following way:

---

[2] Of course the fundamentality of a TP can change over time because it can become more or less fundamentally used. However, the relation in itself does not concern this change.

[3] My discussion is especially tailored on the case of mathematical physics. I do not exclude, however, that it can be suitably extended to other branches of science by taking an appropriately enlarged notion of symbolic practices.

(F) Fundamentality: A TP must be *embedded* in a set of symbolic practices that lead to empirical success.

(R) Robustness: The symbolic practices adopted to deploy the TP must be *reliable*.

Let us begin with (F). This idea hinges on the "embeddedness" of a TP into a set of symbolic practices. An empirical success, a successful prediction or an explanation, is obtained by starting with one TP—or, better, its symbolic codification—and by deriving from it the phenomena to be treated by means of suitable manipulations. In their analysis of the path from TP to success, philosophers usually disregard the epistemic role played by symbolic manipulations of TPs. But if we neglect this important factor of the process of predicting/explaining, we are left with no other option than characterizing fundamentality as a relation between TPs, i.e., a 'Psillosian' criterion and then a 'Lyonsnesque' argument can easily prove that this falls short of providing a satisfactory notion of fundamentality. In my proposal, fundamentality is rather a relation between TP and the symbolic practices adopted to transform and manipulate it. Although intuitively clear enough, the concept of embededdness admittedly needs further philosophical analysis. In Section 4, I provide a historical example to clarify what it means for a TP to be embedded into a set of symbolic practices.

Before discussing the example, however, I need to analyze briefly the idea of robustness. Condition (R) states that reliability, and hence robustness, is a property of the symbolic practices themselves. In other words, and this is the central point, a TP

can be made more robust by means of *historically and rationally describable strategies* conceived to enhance the reliability of symbolic practices adopted to put it to work. One way to appreciate this point is to notice that the concept of reliability has three main components. First, there is an *empirical component*, that is its connection with success. It is expected that reliable symbolic practices have led and will lead to empirical success. This is unsurprising, because it is still part of the relation between (D) and the NMA. Second, there is a *conceptual component*: reliable symbolic practices allow us to distinguish between real facts of nature and artifacts. This is the component that accounts for the non-accidentality of success and it depends on the adoption of strategies to enhance reliability. Applying symbolic practices to multiple cases, relating them with other, better understood, sets of practices (e.g., by showing structure similarities), generalizing solution methods, simplifying computation procedures, introducing redundant check routines, improving the symbolic notation, multiplying proof procedures are just a few examples of strategies used to ensure that the result of symbolic manipulation is a real information and not an artifact generated by the practice itself.[4] Finally, there is a *historical component*. As I said above, deployment is a process extended over time. When are we justified to consider a result as reliable? This is an agent- and a context-dependent component of reliability.

[4] This component of the concept of reliability is closely connected with the usual notion of robustness (see, e.g., (Soler et al. 2012) for an overview). Indeed, robustness has to do with the multiplications of methods of check and control as a way to distinguish what is real and what is fabricated by practices.

I submit that this component can be clarified in terms of *control*. We develop theories because we need to manipulate symbols in order to make predictions and explanations. It is reasonable to state that an agent considers reliable a theory when she has control on it, when she knows how to do things, where the theory can be applied, to what extent, what kind of information she can obtain, what kind of epistemic risks are involved in it, how to improve progressively the performance and a lot of other things related to the general idea of knowing what is going on. Thus, reliability can change over time in virtue of new information and further inquiry. This component accounts for the fact that science is an ongoing human endeavor.

To sum up, I propose to extend SelRealism in the following way:

(SelRealism+) We are entitled to consider the TP *H* as true or approximately true at time *t* if and only if:

1. *H* is embedded into a set of symbolic practices *S*
2. *S* is reliable
3. *H* and *S* lead to significant success

This is a more selective version of SelRealism, because the philosophical and historiographical program stemming from it extends the inquiry to the strategies adopted to improve the reliability of symbolic practices and the contingent conditions for control. As stated in condition 3, the units of analysis of SelRealism+ are TPs-*cum*-

practices rather than TPs only. In the following section, I provide an example of what I mean by intensive and extensive deployment.

4. The Coming of Age of Perturbation Theory

The *Principia Mathematica* are a supreme example of how to embed a TP, in this case the gravitational law, into a set of symbolic practices.[5] However, Newton's mainly geometrical methods were fantastically complicated and notoriously difficult to master. A significant breakthrough in what came to be called celestial mechanics happened in the mid-1740s, when Leonhard Euler laid down the foundations of analytical perturbation theory. Euler made a number of decisive steps forward. First, he used the gravitational law to formulate general equations of motion for celestial problems. Second, he introduced the use of trigonometric series to construct approximate solutions. The use of these series also depended crucially on the gravitational law, because it satisfied the assumption that planetary orbits, even under perturbations, can be represented by a combination of periodic functions. Finally he introduced manipulation practices such as the method of the variation of

[5] In what follows, I consider perturbation theory as the set of practices conceived to put to work the gravitational law. It must be noted that other TPs were involved (e.g., Newton's laws of dynamics) and that the gravitational law can be decomposed in further assumptions such as the action-at-a-distance, the instantaneous propagation and so forth. These considerations affect the level of detail of my example, but not the structure of my argument.

constants and the method of successive approximations to solve the equations of motion. Perturbation theory is therefore a clear example of a set of symbolic practices conceived to cast a TP into a manipulable form and to applied it to specific problems.

For the purpose of this paper, I distinguish two phases in the early history of perturbation theory. The first phase goes roughly from the mid-1740s to the mid-1760s and it concerns the cause of numerous astronomical anomalies. Newton had left behind a few conundrums that even his genius was unable to unravel. The most conspicuous of these problems was the precession of the Lunar apogee. Newton's Lunar theory, elaborated in Book I and III of the *Principia* only managed to obtain half of the observed value. In the 1740s, there were two approaches to the issue of the Lunar apogee. The analytical approach adopted the gravitational law, or a slightly modified form of it, and tried to calculate the observed precession by analytical methods only. The physical approach supposed that the observed anomalies could be due to material causes such as a resisting medium or interplanetary vortices. It is important to realize that these approaches were compatible. Euler himself supported both the resisting medium hypothesis and the analytical approach and occasionally also proposed the use of vortices (letter to Clairaut, 30 September 1747). For several years, the best mathematicians of Europe struggled with the riddle of the Lunar apogee (Bodenmann 2010) until, on 21 January 1749, Alexis Clairaut showed that if one pushes the approximation to the second order of the perturbation, some terms that are negligible at the first order become sizable and generate the missing half of the precession (Clairaut 1752).

Clairaut's success was surely an impressive breakthrough, but what made it so impactful was not the brute fact that gravitational law had eventually led to a successful explanation. Physical hypotheses such as vortices and resisting medium also provided an explanation of the observed precession. The crucial difference lies in the fact that the gravitational law could be fully integrated with the analytical practices and then manipulated to provide suitable symbolic expressions of the precession of the apogee. That did not happen with the physical hypotheses, although not for lack of trying. Euler, for instance, tried hard to integrate the hypothesis of the resisting medium in perturbation theory, but the ensuing equations of motion were simply unmanageable (Euler 1747). Clairaut's success is eminently a story of intensive use of the gravitational law: he managed to integrate it with a set of symbolic practices and to accommodate effectively the observations.

Clairaut's feat did not close the debate on the gravitational law, tough. His calculations used many case-based assumptions, simplifications, and shortcuts and its straightforward extension to more complex cases, such as the behavior of Jupiter and Saturn, was doubtful to say the least. But there was also a deeper problem. At some point in his analysis, Clairaut obtained an "arc of circle", i.e., a trigonometric function multiplied by time. Such terms are obviously unbounded and hence make the whole trigonometric series diverge. Clairaut got rid of it by ad-hoc assumptions, but the status of these unbounded terms remained unclear: they could represent an artifact of the theory, a limitation of its predictive power or even a dynamical instability of the system.

Soon, the problem of the arcs of circle become more troublesome. Euler found the same terms in his analysis of the motion of Jupiter and Saturn and in 1766 Lagrange proved that they are actually a necessary consequence of the method of successive approximations applied to astronomical problems (Lagrange 1766). Thus, in the mid-1760s, perturbation theory appeared to be a fragile set of practices which had scored some important success, but was still marred with problems of unreliability under certain conditions. From the late 1760s onwards, the issue of improving the robustness of perturbation theory became a central preoccupation of the leading mathematicians interested in physical astronomy.

There were two programs inspired by this issue. On the one hand, Lagrange tried to improve the reliability of perturbation methods *as a mathematical theory*. He carried out this project by means of multiple strategies: (1) enhancing the relation between perturbation theory and other branches of mathematics (e.g., potential theory); (2) elaborating arguments to extract information from the equations of motion without solving them (e.g., by using integrals of motion); (3) improving methods to simplify the solution procedure (e.g., Lagrange's coordinates); (4) introducing new symbolic codifications to manipulate the equations of motion (e.g., the perturbing function); (5) making the notation less cumbersome (Lagrange's coefficients). Around the same years, Laplace was also working to improve the reliability of perturbation theory, but his program adopted a different approach. He concentrated on methods to make perturbation theory a more reliable *problem-solving tool*. He developed his own method to eliminate the arcs of circle—which was based on the recalculation of the

integration constants—he imported probability theory and the equations of condition to deal with astronomical observations and devised several strategies to identify in concrete cases those elements of the equations of motion that were likely to produce sizable perturbation terms at higher order. Both Lagrange's and Laplace's programs scored their own successes. In the early 1780s, Lagrange proved a very general result of stability according to which the three more important orbital elements (mean motion, eccentricity, and inclination) are invariable or bounded (Lagrange 1781). Laplace, on his part, explained the decades-long problems of the anomaly in the motion of Jupiter and Saturn as well as the secular acceleration of the Moon (Laplace 1785, 1787; Wilson 1985).

5. Conclusions

In several places, Kyle Stanford has argued that any selection of enduring TPs is ultimately ungrounded and, consequently, the entire SelRealism program is unviable (Stanford 2003, 2006). In his view, there are two possible ways to select essential TPs. The first way is to trust scientists when they say that a certain posit is fundamental. However, neither commonsense, nor, more importantly, historical records support the hypothesis that scientists' take on this matter is or should be particularly reliable. The other option is to wait and see: when a theory is superseded, one can check which TPs have survived. The reason why a selective realist cannot go with this option, however, has been summarized effectively by Peter Vickers:

> If we cannot identify the working posits of a theory until it has been superseded by some other theory, then realism is no longer about identifying what we ought to believe to be true: one is always waiting for the next theory to come along to tell us which parts of our current theory are working posits. (Vickers 2013, 207)

From this, Stanford concludes that SelRealism without prospectively applicable selectivity criteria is empty and should be replaced by a more modest form of realism. But Stanford's wait-and-see stance is neither necessary nor sufficient to do the job it is supposed to do, i.e., to pick out essential TPs. It is not sufficient because there is no guarantee that the TPs survived one theory change will survive the next ones. It is not necessary because we do not need the next theory to form reasonable judgements about essential TPs. As I have shown above, science provides a variety of strategies to improve the reliability of the TP-*cum*-practices and hence good reasons to believe, *within the actual theory*, that a certain TP intensively and extensively deployed is in fact essential.

From this perspective, Stanford's argument simply sets the epistemic bar too high. By stating that the essentiality of a TP can be adjudicated only from the vantage point of the superseding theory, he implicitly challenges the realist to provide a "superselection rule" able to capture the whole history of science, a task that the realist is neither willing, nor actually requested to accomplish. By contrast, the historical and philosophical program of SelRealism+ moves from the conviction that TPs and symbolic practices follow a dynamics able to filter out inessential

components. Consequently, SelRealism+ is committed to historically identify and philosophically analyze this dynamics and to trace the genealogy of our theories in terms of the processes of codification, manipulation, and stabilization of TPs. Ultimately, this program aims at producing new and interesting historical narratives of theory change. It remains true that the strategies making up the theoretical dynamics only provide good reasons to allocate the realist commitment. It might happen that the judgement on the reliability of the TPs-*cum*-practices change over time in virtue of further inquiry or new information. This fact, as stated above, follows from the fallibility of science as a human endeavor and, as such, should not trouble the realist.

Acknowledgements

The research for this paper has been supported by the Marie Sklodowska-Curie Actions, grant no. PIOF-GA-2013-623436.

References

Bodenmann, Siegfried. 2010. "The 18th Century Battle over Lunar Motion." *Physics Today* no. 63:27-32.

Chakravartty, Anja. *Scientific Realism* 2011 [cited 4 February 2015. Available from http://plato.stanford.edu/entries/scientific-realism/.

Clairaut, Alexis. 1752. "De l'orbite de la lune, en ne negligeant pas les quarrés des quantités de meme ordre que les forces perturbatrices." *Memoire de L'Academie Royale des Sciences*:421-440.

Euler, Leonhard. 1747. "Recherches sur le mouvement des corps cèlestes en général." In *Opera Omnia*, 1-44. Leipzig: Teubner.

Harker, David. 2008. "On the Predilections for Predictions." *British Journal for the Philosophy of Science* no. 59:429-453.

———. 2013. "How To Split a Theory: Defending Selective Realism and Convergence without Proximity." *British Journal for the Philosophy of Science* no. 64:79-106.

Kitcher, Philip. 1993. *The Advancement of Science*. Oxford: Oxford University Press.

Lagrange, Joseph Louis. 1766. "Solution de différents problèmes de calcul intégral." In *Œuvres de Lagrange*, edited by Jean A. Serret, 609-668. Paris: Gauthier-Villars.

———. 1781. "Théorie des variations périodiques (Premiére partie contentant les formules générales de ces variations." In *Œuvres de Lagrange*, edited by Jean A. Serret, 347-377. Paris: Gauthier-Villars.

Laplace, Pierre S. 1785. "Théorie de Jupiter et de Saturne." In *Œuvres de Laplace*, 95-239. Paris: Gauthier-Villars.

———. 1787. "Memoire sur les Variations seculaires des Orbites des Planetes." In *Œuvres de Laplace*, 295-306. Paris: Gauthier-Villars.

Laudan, Larry. 1981. "A Confutation of Convergent Realism." *Philosophy of Science* no. 48:19-49.

Lyons, Timothy D. 2006. "Scientific Realism and the Stratagema de Divide et Impera." *British Journal for the Philosophy of Science* no. 57:537-560.

Musgrave, Alan. 1988. "The Ultimate Argument for Scientific Realism." In *Relativism and Realism in Science*, edited by Robert Nola, 229-252. Dordrecht: Kluwer.

Psillos, Stathis. 1999. Scientific Realism: How Science Tracks Truth. London: Routledge.

Soler, Lena, Emiliano Trizio, Thomas Nickles, and William C. Wimsatt. 2012. *Characterizing the Robustness of Science*, *Boston Studies in the Philosophy of Science*. Dordrecht: Springer.

Stanford, P. Kyle. 2003. "No Refuge for Realism: Selective Confirmation and the History of Science." *Philosophy of Science* no. 70 (913-925).

———. 2006. Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives. Oxford: Oxford University Press.

Vickers, Peter. 2013. "A Confrontation of Convergent Realism." *Philosophy of Science* no. 80:189-211.

Wilson, Curtis A. 1985. "The Great Inequality of Jupiter and Saturn: from Kepler to Laplace." *Archive for History of Exact Sciences* no. 33:15-290.

Worrall, John. 1989. "Structural Realism: The Best of Both Worlds?" In *Philosophy of Science*, edited by David Papineau, 139-165. Oxford: Oxford University Press.

Duhemian good sense and agent reliabilism

Famously, according to Duhem a hypothesis can never be experimentally tested in isolation, but only along with the entire theoretical scaffolding it comes with. So in the face of disagreement between theory and experiment, it is impossible to point out which hypotheses in the theory are flawed. A big question for Duhem was, how does the physicist act in such a situation of underdetermination? Which hypotheses does s/he discard, and which one(s) does s/he retain? Duhem's response was that the physicist possesses an intuitive "good sense" that directs this choice. Although good sense does not provide a rigorous, rule-based template for theory choice[1], it allows scientists to weigh evidence and be "fair and impartial" (Duhem, 218) in theory choice.

Recently, there has been much interest in drawing parallels between Duhem's good sense and ideas in virtue epistemology (VE). VE emerged in the 1980s as an approach to epistemology based on virtue ethics. In the words of Greco (2004): "Just as virtue theories in ethics try to understand the normative properties of actions in terms of the normative properties of moral agents, virtue epistemology tries to understand the normative properties of beliefs in terms of the normative properties of cognitive agents." A virtue epistemological reading of good sense as first advanced by David Stump (2007) is based on the idea that Duhem too emphasized the normative properties of the scientist qua cognitive agent and took them as a basis for legitimate scientific

[1] While "theory choice" today is generally understood in the context of contrastive underdetermination, Duhem was primarily concerned with the holist variety of underdetermination and advanced good sense in the context of the latter. But for the purposes of this paper the distinction will not matter, and I shall use "theory choice" to refer to underdetermination in general, as do all the authors I reference.

knowledge in the face of underdetermination of theory by evidence. Stump finds striking similarities particularly between Duhemian good sense and Linda Zagzebski's (1996) views of VE. Here, I discuss the views of Stump, Milena Ivanova (2010), and Abrol Fairweather (2012) in this regard and ultimately propose my own view in response which is an agent-reliabilist reading of Duhem's good sense.

Stump argues that Duhem conceived of good sense in a way that can today be understood as virtue theoretic. In particular, Stump finds similarities between good sense and ideas of VE put forward by Zagzebski (1996). As Stump tells us, Zagzebski argued that justified belief comes from a "cluster of intellectual virtues in the same way that the rightness of an act can be defined in terms of moral virtue in ethical theory"(Stump, 151). Stump argues that Duhem's good sense nicely fits in with these ideas. Good sense depends on the scientist, the cognitive agent, being "virtuous": s/he has to be, in the words of Duhem quoting Claude Bernard, a "faithful and impartial judge". Stump further provides another illuminating quote from Duhem from his lectures on German science:

> "In the realm of every science, but more particularly in the realm of history, the pursuit of the truth not only requires intellectual abilities, but also calls for moral qualities: rectitude, probity, detachment from all interest and all passions. (Duhem, 1991b, p. 43)" (Stump, p. 152).

Stump notes that some of the epistemic virtues put forward by Zagzebski include intellectual sobriety, impartiality and intellectual courage and the list fits very well with Duhem's. Yet another striking similarity between Zagzebski and Duhem according to Stump is that they both appeal to non-rule-governed epistemology. Zagzebski, in making a case for an

epistemology based on ethics, says, "The idea is that there can be no complete set of rules sufficient for giving a determinate answer to the question of what an agent should do in every situation of moral choice." (Stump, 152) Similarly, Duhem arrives at the idea of good sense when the rule-based epistemology of the physical method (i.e. strict agreement between theory and experiment) fails. As Stump says,

> "Holism threatens to make testing impossible, yet Duhem believes that scientific consensus will emerge. While the pure logic of the testing situation leaves theory choice open, good sense does not. Duhem claims that the history of science shows that while there is controversy in science, there is also closure of scientific debates." (Stump, 155)

Milena Ivanova (2010) has argued in response to Stump, that the latter is mistaken in drawing such close parallels between VE and Duhem's good sense. She raises two main objections: first, while VE is concerned with getting to the *truth* via epistemic virtues, for Duhem, physical theory only asymptotically approaches truth – truth here being the truth of a natural order, of the "real affinities" among things. Ivanova makes this point keeping in mind Duhem's view of a 'perfect theory' and the convergent nature of his realism: for Duhem, the aim of physical theory was to classify experimental laws, and a physical theory – one picked out by good sense in the face of underdetermination – constantly approached but never reached, a perfect theory which classified laws and their phenomena in exactly the way underlying metaphysical realities are really classified in nature. So her point is that while VE is concerned with getting to the truth, good sense doesn't help us with that. But as Ivanova herself points out,

> "Still, in response to this objection one can adopt the weaker thesis that even though natural

> classification may not reveal the truth about the unobservable, it will be true for the observable phenomena. Also, one may argue that it is legitimate to aim at a particular epistemic goal independently of whether this goal is achievable or not." (62)

I take her point here to be that both VE and good sense are after all in the business of truth-seeking even though attaining the truth may be impossible for with the latter.

Ivanova's more forceful objection has to do with epistemic justification. According to her whereas VE takes epistemic virtues to be *justifications* for beliefs, Duhem did not invoke the concept of good sense to *justify* belief in one theory over another. (To reiterate, Duhem did not have a full-blown metaphysical notion of truth of a theory – but worked with the surrogate idea of truth, that a right theory approaches a transcendental, natural classification.) Rather, she argues, good sense for Duhem was more a post hoc *explanation* of the physicist's choice: it explains the repeated success of theories at making novel predictions. According to Ivanova, what really justified belief in a theory for Duhem – i.e. the belief that it was approaching a natural classification – was the success of the theory in making correct novel predictions: She says that for Duhem, "[a scientist] is justified in believing that a theory is a natural classification only when some empirical evidence supports it or when the theory has become a 'prophet for us' (Duhem, 27), that is, when it has managed to make novel predictions." (Ivanova, 62). Here's Ivanova's argument broken down:

- Physical theory is a classification of laws.
- In a situation where we have a theory that contradicts experimental data and are left without any means within physics to decide what to do - whether to tweak parts of the theory to accommodate the available experimental data – and if so, which parts to tweak

– or to abandon it for another theory. Somehow in the end, the scientist decides which way to go.

- The "highest test" for physical theory is to ask it to make new and novel experimental predictions.
- When the theory succeeds it is justified – in that it is taken to approach a natural classification.
- Repeatedly, the scientist sees her/his choices made in the difficult situation of underdetermination emerging successful in such predictions.
- How does this happen? There must be some innate ability or virtue in the scientist that enables him to do this: good sense.

Thus according to Ivanova, good sense is an explanation of theory choice rather than a justification for it. Moreover, according to her, Duhem doesn't say anything about good sense as a method of science: he doesn't tell us *how* exactly it directs our choice. His account of how good sense comes about and works to direct theory choice is quite thin. For Ivanova, this further shows that Duhem did not introduce it as a justification but only as a post hoc explanation.

Abrol Fairweather (2012) has argued against Ivanova's above objection and has attempted a position on Duhemian good sense that is a hybrid of Ivanova's and Stump's views. Fairweather claims to draw upon an agent reliabilist VE to do this. Reliabilism in Alvin Goldman's words, "… as a distinctive approach to knowledge is restricted to theories that involve truth-promoting factors above and beyond the truth of the target proposition." (Goldman, 2011) Fairweather's argument is that good sense results in a *reliable* process. Since Duhem's

claim is that good sense has a great "track record" and always picks out a successful theory – i.e. a theory which inevitably *correctly* makes a novel prediction – good sense produces knowledge (which here in the Duhemian context, consists in taking a predictively successful theory to be approaching a natural classification) by a *reliable* process. Good sense is a 'truth-promoting factor' regardless of whether the theory it picks out ultimately succeeds in novel prediction or not. It is "tracking evidentially important features of theories" (Fairweather, 10) Fairweather claims that "If a belief P is the product of a reliable capacity or process this fact constitutes evidence in favor of P." This implies, "If the products of good sense reliably turn out to be supported by compelling new evidence, then being the product of good sense will be evidence for any theory with such a distinguished etiology." (Fairweather, 10) So, Fairweather says, it seems that "future evidence is not required to evidentially distinguish the theory chosen by good sense, because the reliability of good sense is itself evidence supporting that theory." (Fairweather, 10) While I agree that agent reliabilism is the best way to understand good sense, Fairweather does not seem to give an accurate interpretation of this reading. Although he claims to provide an agent reliabilist reading of good sense, he grounds the reliability of good sense in its track record and not in its own nature or the mind where it is born. This is antithetical to agent reliabilist VE which situates reliability in the cognitive character of the agent. So it seems that Fairweather's characterization is more along the lines of process reliabilism or simple reliabilism – according to which a belief is justified just in case it is formed via reliable processes – rather than agent reliabilism, and hence contrary to what he set out to do. His argument does not help situate good sense back into VE. Let us now turn to agent reliabilism in detail.

Greco and Agent Reliabilism: A Short Detour

As above, simple reliabilism is the view that a belief is justified just in case it is formed via reliable processes. Here the proportion of true beliefs the process results in, over time, measures reliability. Greco (1999) argues that simple reliabilism is insufficient for two reasons:

1. An agent might form a belief via fleeting or strange processes: Greco starts by noting that "Reliabilism must somehow restrict the kind of reliable process that is able to ground knowledge, so as to rule out processes that are strange or fleeting." (Greco, 286) As an example of such processes, Greco discusses Platinga's "The case of the epistemically serendipitous lesion" where an agent has a rare kind of a brain lesion, one that makes her believe that she has a brain lesion. There is no evidence for the lesion: there no symptoms, no testimony etc.; in fact there might even be a lot of evidence *against* it. But the agent is unable to take account of this (lack of) evidence due to the lesion. The relevant cognitive process here must no doubt be deemed very reliable, but we would not want to take the resulting belief as justified.
2. Process reliabilism doesn't guarantee that the agent has a subjective justification of her belief. Greco says,

> "[there] is a powerful intuition that knowledge does require that the knower have some kind of sensitivity to the reliability of her evidence. Sometimes this intuition is expressed by insisting that knowledge requires subjective justification. It is not enough that one's belief is formed in a way that is objectively reliable; one's belief must be formed in a way that is subjectively appropriate as well." (285)

Greco's solution to the above problems is agent reliabilism. According to agent reliabilism, reliability is shifted from the belief-forming process to the qualities of the agent's

mind:

> "Relevant to present purposes is Sosa's suggestion for a restriction on reliable cognitive processes; it is those processes that have their bases in the stable and successful dispositions of the believer that are relevant for knowledge and justification. Just as the moral rightness of an action can be understood in terms of the stable dispositions or character of the moral agent, the epistemic rightness of a belief can be understood in terms of the intellectual character of the cognizer." (Greco, 287)

Following Sosa's views, Greco proposes that "knowledge and justified belief are grounded in stable and reliable cognitive character."(Greco, 287) Accordingly, "We may now explicitly revise simple reliabilism as follows: A belief p has positive epistemic status for a person S just in case S's believing p results from stable and reliable dispositions that make up S's cognitive character." (Greco, 287) Hence we see that reliability now has little to do with the truth of the resultant belief(s) but rather with the cognitive character of the agent.

Greco proceeds to show how agent reliabilism also solves the problem of subjective justification:

> VJ: "A belief p is subjectively justified for a person S (in the sense relevant for having knowledge) if and only if S's believing p is grounded in the cognitive dispositions that S manifests when S is thinking conscientiously." (289)

By "thinking conscientiously", Greco clarifies that he does not mean thinking with the purpose of finding truth, but rather the "usual state that people are in as a kind of a default mode – the state of trying to form beliefs accurately." Greco contrasts this with epistemic "vices" such as trying to comfort oneself or trying to seek attention. Lastly, Greco points out that agent reliabilism reverses the "usual direction of analysis between virtuous character and justified

belief". While non virtue theoretic epistemologies understand virtues in terms of justified belief, here justified belief is being cached out in terms of virtues of the cognizer. "Virtuous belief is associated with the dispositions a person manifests when she is sincerely trying to believe what is true", and "The dispositions that a person manifests when she is thinking conscientiously are stable properties of her character, and are therefore in an important sense hers." (Greco, 290) Therefore, a belief formed this way will be subjectively appropriate.

Back to Duhem

Duhem's views seem to exhibit all the features of agent reliabilism discussed above. In addition to the features of good sense and the physicist qua cognitive agent discussed so far I want to draw the reader's attention to Duhem's characterization of the different kinds of minds. For Duhem, the "strong and the narrow" mind is one capable of ordering and organizing laws and hypotheses into theories, and the "supple" mind or the "mind with finesse" – one capable of grasping a wide range of objects and at the same time able to group them logically – is the mind that produces good sense. This certainly seems to talk of "stable dispositions" in Greco's sense of the term, that reflect the "cognitive character" of the scientist. Duhem takes pains to carefully describe the mind of the physicist and discuss beliefs and attitudes *in terms of* cognitive character traits and not the other way round. i.e. Duhem talks of legitimacy of beliefs in terms of cognitive character traits; he does not talk of the traits or "epistemic virtues" so to speak, in terms of the validity of beliefs. For instance, he says about those not interested in seeing a unified system of classification erected, "Only those who affect a hatred of intellectual strength were mistaken to the extent of taking the scaffolding for a completed building." (Duhem, 103) There are several such instances where Duhem turns traditional non virtue-theoretic epistemology on its head and makes cognitive character traits basic. Now it remains to be seen if we can defend a view of

justification from good sense that goes with Greco's account. If we are successful in this, Ivanova's position will be untenable. Before going there though, let us return to Fairweather for a moment.

In addition to the argument from reliabilism, Fairweather advances another argument against Ivanova's "deflation of good sense": the position that good sense does not lend any epistemic strength or any justification to the chosen theory. The argument is that if good sense were indeed merely explanatory and post hoc as Ivanova claims, and not justificatory, then we are free to imagine a case where good sense doesn't intervene at all. After all, if good sense explains theory choice and there is no choice being made – i.e. no explanandum - we don't need an explanation. So let us suppose that we don't make any choice and just wait for a future novel prediction to make a choice and justify it. This might not be the most efficient way to choose a theory, but let us assume we do this nevertheless – for according to Fairweather, Ivanova's objection should imply the possibility of this solution. Fairweather rightly points out that in this situation we *might* again end up with an underdetermination: what if all competing theories pass the novel prediction test? Therefore, Fairweather argues, good sense must play an important epistemic role above mere explanation, in the face of such a "second level" underdetermiantion. But he goes further than that and says that without it, we *would* never end up with a determinate choice, even with new confirming evidence. What Fairweather is ignoring here is that future evidence *could* pick out a theory, however small the probability. It is possible that when all the options resulting from underdetermination are asked to make a novel prediction, only one succeeds, hence obviating the need for any further theory revision. But the important point is that good sense enters the scene even before such an attempt to single out a theory based on novel prediction. So the merit of good sense in my view does not lie in the inability of novel

predictions to single out a theory. It is more fundamental than that. But reasons for meriting good sense apart, let us again look at Fairweather's take on *what* the merit of good sense is.

According to Fairweather, good sense confers *uniqueness* to a theory (which, according to him, no future evidence can confer). But after good sense has uniquely picked out a theory, it is a successful novel prediction that counts as evidence in favor of the chosen theory. Fairweather makes the following interesting observation that follows from such a reading of good sense:

> "This shows an interesting fact that new evidence in favor of a theory gives it a different epistemic standing depending on whether we are considering it alongside or independent of meaningful rivals. In the former case, new confirming evidence does not make a theory the determinate choice with fundamental epistemic standing. In the latter case, that same evidence determines theory choice and confers fundamental epistemic standing." (Fairweather, 13)

So there are two "epistemic values and epistemic standings": uniqueness, which comes from good sense, and clinching evidential support from a successful novel prediction. This way, good sense alone does not confer "fundamental epistemic standing", and evidence alone cannot confer uniqueness. This account which recognizes an important epistemic role for both good sense and new evidence, Fairweather calls the "hybrid reading".

My own view is that while Fairweather is right in that good sense plays a key epistemic role unlike what Ivanova says, we can go back full circle to Stump and have a proper virtue epistemological – specifically agent reliabilist – reading of good sense. I contend that good sense confers not just uniqueness, but actually does determine theory choice, also providing (an agent-

reliabilist) justification. Good sense doesn't simply pick one and put the rest "out of the running". It is not just something that prevents the proliferation of acceptable theories obtained by tweaking different parts of theories that don't agree with future experiment. Good sense provides a *basis* for the uniqueness. Just as with the problem of coming up with a realist interpretation of Duhem, this problem of the epistemic role of good sense is not easy either given the sometimes confusing nature of Duhem's claims. Nonetheless, I still think an agent-reliabilist VE reading of Duhem is possible and that Ivanova and Fairweather are mistaken.

Ivanova claims that good sense is only offered as a post hoc explanation of theory choice during underdetermination and not as a justification. I argue to the contrary. Ivanova's claim seems to be based on a purely externalist notion of justification. It seems to assume that there is one single concept of justification – specifically, externalist, evidential – and that good sense doesn't fit with it. But justification can be of many kinds. Duhem says we can "very properly decide" (Duhem, 217) between multiple theory choices using good sense. Further, he says good sense strongly "comes out in favor of" one of the choices – again implying that we are compelled to accept its judgment *even before* future experiment can ratify the choice. He goes on to say, "Pure logic is not the only rule for our judgments; certain opinions which do not fall under the hammer of contradiction are in any case perfectly unreasonable." (Duhem, 217) How do we understand such language? If an epistemic choice is proper, forceful, and reasonable, I don't see any reason we cannot properly construe it as being justified, in an *internalist* sense.

Further, Duhem does *not* introduce good sense as a merely post hoc explanation. He says, we can "properly decide" between the various options of theories using good sense. "Properly

decide" very much implies an active role for good sense *during* underdetermination. Duhem presents elaborate and careful characterizations of different kinds of minds and puts forward quite clearly, *normative* merits of cultivating/ possessing one kind of mind over the other as far as physics goes (the supple or the strong and narrow over the ample, broad and weak). Good sense is but a feature of the supple mind. It is not introduced all of a sudden as a new idea to just "save the (meta)phenomenon" of theory choice during underdetermination. It is a smooth and natural continuation of Duhem's views on the mind of the theorist, which he articulates way before he comes to this problem of underdetermination, in one of the early chapters in *Aim and Structure*. In fact, Duhem's view that physicists don't actually actively choose hypotheses at all, and that they "come to his mind" when his mind is ready to receive them, clearly reveals the agent reliabilist in Duhem.

Finally, Greco's account of agent reliabilist justification seems to lend itself to Duhem very well. Reliable cognitive character *justifies* beliefs it produces and further, it is subjectively justified: Duhem's virtuous scientist certainly "thinks conscientiously", following Duhem's instructions of shunning passions and interests, and so a belief, here the belief in the theory chosen, grounded in the cognitive dispositions, here good sense, he manifests when thinking like this – is subjectively justified. So we seem to have comfortably accommodated Duhem in a full-blown agent reliabilist reading.

But what about the textual evidence cited by Ivanova, which seems to say Duhem did not think good sense justified theory choice? Why does Duhem insist that despite good sense, it is a successful novel prediction that has the final word? Why does he, in the context of resolving

underdetermination say in as many words that the method of the physicist "is justified only by experiment"? I contend that throughout *Aim and Structure*, Duhem seems to have two distinct, non-intersecting epistemologies: one of physics, and one outside of physics – which we may call philosophy. Duhem was a physicist-philosopher. He frequently claims that although there are absolutely no epistemic resources *within* physics for us to believe that physical theory latches on to a natural underlying order, we are forced to believe so by various factors outside of physics, logic and reason. It is worth noting that Duhem cites Pascal as saying that we sometimes believe for 'reasons that reason does not know', both in the context of theories converging on to a natural classification as well as in that of good sense during underdetermination. About the former, he says: "The opinion is a legitimate one because it results from an innate feeling of ours which we cannot justify by purely logical considerations, but which we cannot stifle completely either." (Duhem, 102) Further:

> "No language is precise enough and flexible enough to define and formulate them; and yet, the truths which this common sense reveals are so clear and so certain that we cannot either mistake them or cast doubt on them; furthermore, all scientific clarity and certainty are a reflection of the clarity and an extension of the certainty of these common-sense truths." (Duhem, 104)

Since Duhem attributes good sense to similar patterns of thinking, we can associate his above assertions about the legitimacy of beliefs not borne out of logic, with good sense as well. Given Duhem's commitment to the moral goodness and the intellectual acuity of the supple, strong and narrow minds, it is very unlikely that he would think that epistemic ends justify the means (here, successful novel prediction justifying that which chose the theory, i.e. good sense). Reliabilism in fact expressly turns this around and say it is the means (by virtue of their

reliability) that justify the ends. So beliefs that arise from good sense are *justified* from an (internalist, deontological) agent reliabilist perspective. The justification Duhem talks about when he says that the methods of the physicist are justified by experiment should be when we are strictly within the context of physics: there it is Duhem qua physicist speaking. But from a broader, philosophical perspective, Duhem rather means, I think, that experiment *validates* the choice and confers *certainty* on it. But we can have justification without certainty, like in agent reliabilism. In simpler terms, the *reasons* for which the physicist chooses a theory are grounded in her good sense. However, the successful novel prediction will no doubt make the choice certain.

Thus, Ivanova is mistaken in arguing that good sense does not provide justification. Fairweather's hybrid reading is inadequate as well for it ignores the justification offered by a proper agent reliabilist reading of good sense. I argue that a proper agent reliabilism accommodates Duhem as a virtue epistemologist very well and shows us that good sense does offer justification for theory choice. Importantly, I have shown that it is certainly not a post hoc explanation but a part and parcel of Duhem's overall views on the mind of the physicist.

References

- Duhem, Pierre. (1954). *The aim and structure of physical theory*. Princeton: Princeton University Press.
- Fairweather, A. (2012) 'The Epistemic Value of Good Sense' *Studies in the History and Philosophy of Science* http://philpapers.org/archive/FAITEV.pdf

- Goldman, Alvin. 'Reliabilism', *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2011/entries/reliabilism/>.

- Greco, J. 1999. 'Agent reliabilism' in *Philosophical Perspectives* 13: 273-296.

- Ivanova, M. (2010). 'Pierre Duhem's good sense as a guide to theory choice'. *Studies in History and Philosophy of Science*, 41, 58–64.

- Stump, David. (2007). Pierre Duhem's virtue epistemology. *Studies in History and Philosophy of Science, 38, 149–159*.

There *Is* a Special Problem of Scientific Representation

(Word count: 4998)

Abstract: Callender and Cohen (2006) argue that there is no need for a special account of the constitution of scientific representation. I argue that scientific representation is communal and therefore deeply tied to the practice in which it is embedded. The communal nature is accounted for by *licensing*, the activities of scientific practice by which scientists establish a representation. A case study of the Lotka-Volterra model reveals how the licensure is a constitutive element of the representational relationship. Thus, any account of the constitution of scientific representation must account for licensing, meaning that there *is* a special problem of scientific representation.

1. Introduction

According to many philosophers of science, representation in scientific practice is different from representation in other disciplines, like art and language. This claim is denied by Craig Callender and Jonathan Cohen (2006), who argue that representation is the same across disciplines. In this paper, I will argue that their view leaves the communal nature of scientific representation unexplained. To explain why scientific representation is dependent upon practice, I will introduce the concept of licensing, in which the targets of representational vehicles are determined through various activities performed by scientists in accord with broader scientific practice. I will argue that licensure is a constitutive feature of representation in science, indicating that there *is* a special problem of scientific representation.

2. Callender and Cohen's View

On Callender and Cohen's evaluation, much of the literature on scientific representation has been "concerned with non-issues" (2006, 67). Specifically, they think there is no reason for philosophers of science to give a special account of the "constitution question:" "What constitutes the representational relation between a model and the world?" (2006, 68). In response to this question, they make a few observations. One is that it is "economical and natural to explain some types of representation in terms of other, more basic types of representation" (2006, 70). They also identify a general desire to have a consistent account of how "entities other than models—language, pictures, mental states, and so on—…represent the very same targets that models represent" (2006, 71). For these reasons, they suggest that

"scientific representation is just one more special case of derivative representation" (2006, 75). That is to say that the representational nature of scientific vehicles is explained in the same way that the representational nature of linguistic entities, artwork, etc. is explained. In each case, and in every practice, the representational nature in question will be reduced to a more fundamental representational entity. So, e.g., the representational nature of a word, a painting, and a scientific model will each be explained in terms of the representational nature of mental states.

On Callender and Cohen's view, representation is purely stipulative: "virtually anything can be stipulated to be a representational vehicle for the representation of virtually anything…" (2006, 74). Of course, it is not the case that *any* stipulated representation will actually be useful for scientific aims. Thus, they identify pragmatic constraints which delimit scientific representation. However, they make it quite clear that these constraints are delimiting *already-existing* representations. As such, the pragmatic constraints are not a part of an account of the constitution of representation itself: "the questions about the utility of these representational vehicles are questions about the pragmatics of things that are representational vehicles, not questions about their representational status per se" (2006, 75).

If Callender and Cohen are correct, then we are left rethinking a rather extensive literature on scientific representation which typically begins with the assumption that there *is*

something special about representation in science.[1] As one example among many, Mauricio Suárez (2004) defends an inferential conception of scientific representation. His account takes careful notice of the aims of scientific practice, noting that mere stipulation (what he calls "representational force") is insufficient for representation in science. To be a *scientific* representation, a vehicle must also permit surrogate reasoning which "allows competent and informed agents to draw specific inferences regarding [a target]" (2004, 773). If we accept Callender and Cohen's view, then Suárez's account and the many others like it do nothing more than identify some of the typical pragmatic strategies employed in delimiting representations for scientific uses (Callender and Cohen 2006, 78).

3. Private Reminiscence and Communal Representation

In order to show that the extensive literature on scientific representation has not been addressing a non-issue, I will need to show that there is a special problem of scientific representation, a feature unexplained by Callender and Cohen's account. I submit that the relevant feature in need of special explanation is the communal nature of scientific representation, that it inherently involves reference to the practice. To see why Callender and

[1] For more accounts which answer the constitution question in a distinct way, see the work of Ronald Giere (1988, 2004), Bas van Fraassen (1980, 2008), RIG Hughes (1997), Steven French, James Ladyman, and Otávio Bueno (French and Ladyman 1999; Bueno and French 2011), and Gabriele Contessa (2007). For an overview of these accounts of scientific representation among others, see Brandon Boesch (2015) and Mauricio Suárez (2015).

Cohen's view is unable to account for the communal nature of scientific representation, consider what I call 'reminiscence', a representational relationship which lacks the same communal feature. It is defined schematically as the following:[2]

> Some X is reminiscent of some Y for some agent A provided that when A thinks about or experiences X, she thinks about or experiences Y and attributes some connection between X and Y.

So, for example, a drawing can be reminiscent of my nephew, the smell of honeysuckle can be reminiscent of golfing, etc.

There are three noteworthy features of reminiscence. First, the representational nature of reminiscence can be reduced to the representational nature of more fundamental entities. For example, I can explain the drawing's reminiscence of my nephew in virtue of the mental state produced by the drawing (which is about my nephew, who created it). Second, stipulation is sufficient to create an instance of reminiscence. For example, I could draw a symbol on my hand which I create for the sake of reminding me to buy bread from the store. The reminiscent relationship exists because of my stipulative act. Finally, any limitations of reminiscent relationships will be made for pragmatic reasons. For example, it would be for pragmatic reasons that I make the symbol on my hand look like a loaf of bread.

[2] I should note that the account of reminiscence here is not meant as a detailed explanation of this concept, but only as an analogy to draw a point about representation.

These three features of reminiscence are noteworthy because they are shared by Callender and Cohen's view of scientific representation. In fact, from Callender and Cohen's perspective, the only major difference between the two concepts would be the particular aims for which each relationship is utilized. While important, these different aims alone are insufficient to explain a key dissimilarity between scientific representation and reminiscence: while reminiscence can be private, scientific representation is necessarily communal. That reminiscence can be private can be seen from the fact that discussions of reminiscence can terminate in disagreement. For example, no one is ultimately 'correct' about whether or not someone is reminiscent of someone else. This is because reminiscence is agent-relative and so depends only upon some particular agent and her mental states.

Scientific representation relies on much more. As Suárez has argued, "representation is not at all 'in the mind' of any particular agent. It is rather 'in the world', and more particularly in the social world – as a prominent activity or set of activities carried out by those communities of inquirers involved in the practice of scientific modelling" (2010, 99). Scientific representation is not isolated from the practice in which it is embedded. It is necessarily communal.[3] The communal nature is demonstrated from the fact that representational vehicles demonstrate autonomy from individual scientists and their mental

[3] The view of representation argued for in this paper echoes many of the points made by Ludwig Wittgenstein's in his 'Private Language Argument' where he argues that meaning is necessarily communal (1953/2009, 95^{e}-111^{e}).

states.[4] For example, a scientist's rogue stipulation that the Lotka-Volterra model (which represents predator-prey relations) represents population change due to genetic drift does not count as an instance of scientific representation. This is not only because it does not (pragmatically) allow for meaningful insights, but also because it ignores and discounts the autonomous elements of the model as understood by the broader scientific community.[5] The autonomous elements are seen in the materiality or historicity of the representational vehicle; in its development, reception, and contemporary use. Understanding how and why the scientific object represents its target requires paying attention to these communal features. That is to say that the communal nature is partially *constitutive* of the representational relationship. Callender and Cohen's account of scientific representation does not sufficiently account for these constitutive communal elements, as will be shown more explicitly below.

4. Licensing

Explaining the communal nature of scientific representation requires that attention be given to the material, autonomous dimensions of the representational vehicle in terms of its

[4] This point has already been made specifically with regard to models by Morrison and Morgan (1999). Here, I am extending a similar point to other representational vehicles, including things like diagrams and figures.

[5] Of course, there may be disagreements and developments internal to the practice about how to use some representation, but these disagreements and developments are *part of the practice*.

development, reception, and use. All of these features partially establish a scientific representation, through an activity I call *licensing*. Licensing is the set of activities of scientific practice by which scientists establish the representational relationship between a vehicle and its target. It is itself a constitutive element of the representational relationship: it is a critical part in explaining how and why some vehicle represents its target. Seeing the sorts of activities involved in licensing and how they partially constitute the representational relationship will require that we pay close attention to the historical development, reception, and use of actual instances of scientific representation.

4.1 Licensing in Artistic Representation

A similar sort of licensing is present in representation in art, and so an initial pass on the concept as it applies to artistic practice will be helpful to draw an analogy to licensing in science.[6] To see the role of licensing in artistic representation, consider an example. The mere stipulation that Pablo Picasso's *Guernica* should represent the pain of cyberbullying is clearly insufficient to make it represent this target. Understanding how *Guernica* is representational involves an awareness of communal features: Picasso's intentions within the environment in which he created the painting, how the painting was received by viewers in the years following its creation, and how it is understood today. With these features in mind,

[6] It is somewhat contentious to draw conclusions about the nature of representation in science by appeal to art; see e.g. Bueno and French (2011). Nonetheless, it is a common technique in discussions of scientific representation; see e.g. Suárez (2004).

it is clear that *Guernica* represents the pain and suffering of the people of Guernica who had been bombed by axis forces at the request of Francisco Franco and the Spanish Nationalists. The licensing here is a constitutive element of *Guernica*'s representational nature: without these features, it is not clear whether or how the painting would manage to represent anything at all.

Licensing also occurs outside of the scope of authorial intent, when the artistic community comes to accept that a piece of art is representational in a way that was not intended by the author. A good example can be taken from an anecdote related by the author Flannery O'Connor:

> [A] student asked me…: "Miss O'Connor, what is the significance of the Misfit's hat?" Of course, I had no idea the Misfit's hat was significant, but finally I managed to say, "Its significance is to cover his head." (1988, 853)

The Misfit is a key character in O'Connor's famous short story, "A Good Man is Hard to Find," and, as such, it would not be surprising for his wardrobe to be importantly representational. Her answer indicates that while she did not intend any representational target for the hat, there may yet be one. If the hat is representational, it will not be due to her authorial intent, but rather due to the views of the broader artistic community.

Let me make it very clear that the licensure so far described is not already accounted for by elements of Callender and Cohen's account. First, notice that none of these means of licensing is a mere pragmatic limitation of already existing representations. It is not as if *Guernica* represents anything and everything, but is then *limited* by the contexts of Picasso,

audiences, and art historians. These contexts are a crucial part of understanding why it represents at all. Nor is the licensing mere stipulation. O'Connor leaves it open that there may be a representational target for the Misfit's hat, even though she did not stipulate one. A single reader's stipulation alone is insufficient to make it a representation, since the target must also fit well with the Misfit's characteristics, with O'Connor's general themes as understood by literary critics and audiences alike, and so on. Once again, these contexts are a critical part of establishing the representational nature of the hat.

4.2 Licensing in Scientific Representation: A Case Study

The unique aims of science indicate that the licensing of scientific representation is of a different kind than the licensing in art. All the same, licensing similarly plays a critical role in establishing scientific representation. According to Tarja Knuuttila, case studies of scientific representation have revealed that it is "a complicated phenomenon" and "a laborious art" (2014, 304). Understanding the nature of licensing and its role in the complexities of scientific representation will be best accomplished by examining the complicated features seen in the context of a case study. Examples could be made of any type of representational vehicle, like the masterful case study of a scientific figure made by Bruno Latour (1999). I will take as my example the Lotka-Volterra model, since its development exhibits interesting features, many of which have already been widely discussed by other philosophers (e.g. Knuuttila and Loettgers 2011, forthcoming).

As mentioned above, the Lotka-Volterra model is used by ecologists to represent predator-prey relations. It had its beginnings in the independent work of two different

scientists, Vito Volterra and Alfred Lotka. In understanding the representational nature of this model, it is important to pay attention to the licensing through its historical development. This attention includes noticing things like the way that the construction of the model by Lotka, Volterra, and others has been responsive to certain theoretical and empirical aims. These historical and practice-centered features of the model's development reveal the partial autonomy of its representational nature. These features constitute the licensing which is itself partially constitutive of the representational nature of the model since understanding how and why the model represents its targets requires attending to these features. Let us now turn to examine these features in more detail.

Consider first the development of the model by Volterra, who was "motivated by the goal of reproducing the kind of oscillating behavior that was observed empirically in fishery statistics" (Knuuttila and Loettgers forthcoming, 19). His aim to address a theoretical question with an empirically useful model is central not only to understanding how the model historically came about, but in understanding how it represents its targets. Consider how Volterra described his project and the aims which permeate his description:

> Let us seek to express in words the way the phenomenon proceeds roughly: afterwards let us translate these words into mathematical language. This leads to the formulation of differential equations. If then we allow ourselves to be guided by the methods of analysis we are led much farther than the language and ordinary reasoning would be able to carry us and can formulate precise mathematical laws. These do not contradict the results of observation. Rather

> the most important of these seems in perfect accord with the statistical results. (1928, 5)

Volterra's actual process of moving from words, to equation, to application of results (for both theoretical and empirical purposes) first involved creating an equation to account for the population change of a single species. He then added additional species and modelled interactions under different conditions, including, notably, contending for the same food and the predation of one species upon the other. Using these models, he demonstrated "three fundamental laws of the fluctuations of the two species living together" (1928, 20). He then applied these theoretical laws of predator-prey relations to the empirical case which had prompted his analysis, the peculiar rise in predator populations during the decrease of fishing of prey populations in the Adriatic Sea during World War I (1928, 21).

Why does Volterra's model represent these theoretical features of predator-prey relations? Why does it represent the populations of fish in the Adriatic during World War I? It represents these targets because, through a series of steps of analysis, revision, and development, each of which was responsive to certain theoretical and empirical aims understood and described in his account, Volterra *established* this representational nature. Indeed, as explained by Knuuttila and Loettgers (forthcoming), the historical development of this model has a much more extended history than the one Volterra described in the two papers where he first introduced it (1926, 1928). The model is a representation of its target not by mere stipulation and pragmatic constraint, but through careful and attentive construction of equations which ensure that the model functions in the wider theoretical

contexts and can explain the relevant empirical aims. In short, the model represents its targets because Volterra so *licensed* it by building into the model these external, autonomous representational features. Without these features, how or what would it represent?

Consider another instance of licensing in the development of the Lotka-Volterra model, this time by Lotka. His development proceeded with a different aim than Volterra: "instead of starting from the different simple cases and generalizing from them, he developed a highly abstract and general model template that could be applied in modelling various kinds of systems" (Knuuttila and Loettgers forthcoming, 13). He began by creating a very general equation which described "evolution as a process of redistribution of matter among the several components…of the system" (Knuuttila and Loettgers forthcoming, 15). In two papers (1920a, 1920b), Lotka applied this general equation to particular cases in biology and chemistry, in each case coming to theoretical conclusions about the systems in question. For example, in applying the equation to a predator-prey system, he concluded that there would be "undamped oscillation continuing indefinitely" among the two populations (1920a, 414). Lotka did not specifically apply the results to any empirical data, but instead used his results to come to theoretical conclusions about these relationships which he then connected to theoretical ecological principles drawn from Herbert Spencer's *First Principles* (1920a, 414).

Why does Lotka's model represent its theoretical target? What constitutes this representational relationship? Any attempt to explain the representational relationship must reference the way in which Lotka derived his general equation and the way in which he applies it to the specific cases. That is to say, the representational nature of the model is

constructed through the scientific activities performed by Lotka during the development of the model. Lotka does not merely stipulate that his model targets predator-prey relationships. Instead, he builds this ability into the model during the development of the general equation and further constructs this ability in his application of the question to specific targets. In so doing, he partially constructs the representational nature of the model—he licenses it as a representation through activities in accord with the broader practice.

The Lotka-Volterra model's history since its initial development is long and complex. As described by Alan Berryman (1992), one development was a shift in the 1940s to the use of a logistic formulation which allowed for attention to be placed on predator-prey ratios rather than products. Another development, which occurred around the same time, was the use of a predator functional response which introduced a nonlinear rate of death for the prey. These developments license new representational targets by expanding and altering the model to make it responsive to different theoretical or empirical aims, by removing idealizations, or otherwise by allowing for different theoretical conclusions. Many other variations of the Lotka-Volterra model exist, licensed by similar developments. Additionally, the original formulation of the model is still used in introductory textbooks on ecology (see, e.g. Cain, Bowman, and Hacker 2008). The representational nature of the model in each of these cases is partially established by these features of the model which stand independent of any mental states of scientists and students alike. In short, the constitution of the representational nature of the Lotka-Volterra model relies deeply upon these historical features of licensing as understood by the broader scientific community.

Let me briefly underscore the importance of these activities of licensing to the representational nature of the Lotka-Volterra model by imagining a scenario in which these features are absent. Suppose that Volterra and Lotka had proceeded differently. Suppose that they began, for no particular reason, by drawing a five-pointed star and stipulated that it represented predator-prey relations. What is the status of this star, qua representation? It is not as if the star *really* is a scientific representation of predator-prey relations albeit a bad representation (because it does a poor job of meeting certain pragmatic constraints). Rather, the star plainly fails to be a scientific representation at all. Scientific representations are constructed to assist in answering certain questions, explaining certain phenomena, understanding certain target systems. It is through licensing that scientists build into the vehicle the features capable of achieving these aims. A vehicle without licensing does not have this ability and so it is not just a bad representation. It is not a representation *at all.* Indeed, a discussion of the representational nature of vehicles which lack these features is either infelicitous or involves an equivocation of the word 'representation.' A view of scientific representation which equally counts both the star and the Lotka-Volterra model as full scientific representations, even if it specifies one as good and one as bad, underestimates the role of these historical features of the model. They are not external to the representational nature of the vehicle, but are themselves an essential constitutive feature of this representational nature: without these features, the vehicle is not a scientific representation at all.

5. The Special Problem of Scientific Representation

If I am right that licensing is a necessary constitutive feature of scientific representation which explains its communal nature, then contrary to Callender and Cohen's suggestion, we cannot pull the question of the constitution of representation away from questions of practice. A scientific object represents its target not (only) because there is some stipulation and pragmatic constraint, but also in virtue of licensing: the context in which it was created, the application of theoretical and empirical constraints, the awareness of and management of idealizations, and the history of its reception and use. Accounting for whether and how a scientific object represents its target will always require reference to these features which partially establish the representational nature. Thus, there *is* a special problem of scientific representation.

I should note that I am not here arguing for a stronger counter claim to Callender and Cohen which says that accounts of the representational nature of mental states are without *any* value to the constitution question of scientific representation. But my argument does indicate that an account of the representational nature of mental states *alone* is insufficient to account for scientific representation. Even if tomorrow we had a solid, universally accepted account of the representational nature of mental states, we would not yet have a complete account of scientific representation. We would still need an account of the deep reliance that it has upon the practice in which it is embedded. Thus, while our discussion of the constitution of scientific representation might include reference to the representational nature of mental states, it must also include reference to what I have described here as the licensing by the practice.

A different concern is that the use of the word 'special' is a bit deceptive. What I have identified here as the 'special' problem of scientific representation turns out to be a common feature of representation across disciplines, since, for example, I have suggested that it holds of artistic representation as well. While it is true that, according to my argument, an account of artistic representation will likely take account of licensing as well, it does not indicate that it is the *same type* of licensing in both practices. Indeed, given the unique aims that mark off scientific practice, its licensing can reasonably be expected to be correspondingly unique. That is to say that understanding, knowing, or explaining the empirical world are special aims, and therefore subject to special sorts of licensing. Scientific representation remains special because these features merit special attention.

We might also wonder whether it is right to continue to discuss scientific representation as a whole. If understanding representation in science requires in part that we understand the way in which scientists of a practice develop, utilize, and adapt these representational devices, then it is at least possible that these activities will be different within different domains. For example, the licensure of representations in physics might be rather different from that of economics. My suspicion is that, given the common broad scale aims of the various domains, we can still say some general things about representation in science as a whole. Nonetheless, we would do well to pay attention to representation as it occurs in these more localized contexts. Moving forward from this conclusion to develop further insights about the nature of scientific representation will involve analyzing specific representational objects or strategies as they occur in scientific practice, perhaps taking hints and clues from

in-the-field investigations like those conducted by sociologists of science, e.g. those in Lynch and Woolgar (1990), Latour (1999), and Coopmans et al. (2014).

6. Conclusion

Though Callender and Cohen's view remains a formidable approach to the constitution question of scientific representation, I have endeavored in this paper to show why their account is insufficient, and thus why this question merits continued attention by philosophers of science. Representation in science is deeply tied up with the practice in which it is embedded. The communal nature of scientific representation can be seen in the way that science, as a practice, partially constructs its representations through the activities of licensing. The licensing is not the pragmatic limitation of some already existing representations, but is itself a constitutive element of the representational relationship. Any account of what it is for a scientific object to represent its target will necessarily involve reference to licensing. Thus, there *is* a special problem of scientific representation.

Bibliography

Berryman, Alan. 1992. "The Origins and Evolution of Predator-Prey Theory." *Ecology* 73: 1530-1535.

Boesch, Brandon. 2015. "Scientific Representation." *Internet Encyclopedia of Philosophy*. http://www.iep.utm.edu/sci-repr/

Bueno, Otávio, and Steven French. 2011. "How Theories Represent." *British Journal for the Philosophy of Science* 62: 857-894

Cain, Michael, William Bowman, and Sally Hacker. 2008. *Ecology*. Sunderland, MA: Sinauer Associates, Inc.

Callender, Craig, and Jonathan Cohen. 2006. "There Is No Special Problem About Scientific Representation." *Theoria* 21: 67-85.

Contessa, Gabriele. 2007. "Scientific Representation, Interpretation, and Surrogative Reasoning." *Philosophy of Science* 74: 48-68.

Coopmans, Catelijne, Janet Vertesi, Michael E. Lynch, Steve Woolgar (eds.). 2014. *Representation in Scientific Practice Revisited*. Cambridge, MA: MIT Press.

French, Steven and James Ladyman. 1999. "Reinflating the Semantic Approach." *International Studies in the Philosophy of Science* 13: 103-119.

Giere, Ronald. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

———. 2004. "How Models Are Used to Represent Reality." *Philosophy of Science* 71: 742-752.

Hughes, R.I.G. 1997. "Models and Representation." *Philosophy of Science* 64 (Proceedings): S325-S336.

Knuuttila, Tarja, and Andrea Loettgers. 2011. "The Productive Tension: Mechanisms Vs. Templates in Modeling the Phenomenon." In *Models, Simulations, and Representations*, ed. P. Humphreys and C. Imbert, 3-24. New York: Routledge.

———. Forthcoming. "Modelling as Indirect Representation? The Lotka-Volterra Model Revisited." *British Journal of Philosophy of Science*, in press.

Knuuttila, Tarja. 2014. "Reflexivity, Representation, and the Possibility of Constructivist Realism." In *New Directions in the Philosophy of Science,* ed. M. C. Galavotti, S. Hartmann, M. Weber, W. Gonzalez, D. Dieks, and T. Uebel, 297-312. Dordrecht, The Netherlands: Springer.

Latour, Bruno. 1999. Circulating Reference. In *Pandora's Hope.* Cambridge: Harvard University Press.

Lotka, Alfred. 1920a. "Analytical Note on Certain Rhythmic Relations in Organic Systems." *Proceedings of the National Academy of Arts and Sciences* 42: 410-415.

———. 1920b. "Undamped Oscillations Derived from the Law of Mass Action." *Journal of the American Chemical Society* 42: 1595-1598.

Lynch, Michael E., and Steve Woolgar (eds.). 1990. *Representation in Scientific Practice.* Cambridge: MIT Press.

Morgan, Mary, and Margaret Morrison (eds.). 1999. *Models as Mediators: Perspectives on Natural and Social Science.* New York: Cambridge University Press.

O'Connor, Flannery. 1988. "The Catholic Novelist in the Protestant South." In *Flannery O'Connor: Collected Works,* 853-864. New York: Literary Classics of the United States.

Suárez, Mauricio. 2004. "An Inferential Conception of Scientific Representation." *Philosophy of Science* 71: 767-779.

———. 2010. "Scientific Representation." *Philosophy Compass* 5: 91-101.

———. 2015. "Representation in Science." In *The Oxford Handbook of Philosophy of Science*, ed. P. Humphreys. New York: Oxford.

van Fraassen, Bas C. 1980. *The Scientific Image*. New York: Oxford University Press.

———. 2008. *Scientific Representation: Paradoxes of Perspective*. New York: Oxford University Press.

Volterra, Vito. 1926. "Fluctuations in the Abundance of a Species Considered Mathematically." *Nature* 128: 558-560.

———. 1928. "Variations and Fluctuations of the Number of Individuals in Animal Species Living Together." *Journal du Counseil International Pour l'Exploration de la Mer* 3:3-51.

Wittgenstein, Ludwig. 1953/2009. *Philosophical Investigations,* trans. G.E.M. Anscombe, P. M. S. Hacker, and J. Schulte. Malden, MA: Wiley-Blackwell.

Dissolving the missing heritability problem

Abstract: Heritability estimates obtained in genome-wide association studies (GWAS) are much lower than those of traditional quantitative methods. This has been called the "missing heritability problem". By analyzing and comparing these two kinds of methods, we first show that the estimates obtained by traditional methods involve some terms that GWAS do not. Second, the estimates obtained by GWAS do not take into account epigenetic factors transmitted across generations, whilst they are included in the estimates of traditional quantitative methods. Once these two factors are taken into account, we show that the missing heritability problem can be largely dissolved. Finally, we briefly contextualize our analysis within a current discussion on how non-additive factors relate to the heritability estimates in GWAS.

1. Introduction.

One pervasive problem encountered when estimating the heritability of quantitative traits is that the estimates obtained from Genome-Wide Association Studies (GWAS) are much smaller than that calculated by traditional quantitative methods. This problem has been called the missing heritability problem (Turkheimer 2011). Take human height for example. Traditional quantitative methods deliver a heritability estimate of about 0.8, while the first estimates using GWAS were 0.05 (Maher 2008). More recent GWAS methods have revised this number and estimate the heritability of height to be at most 0.45 (Yang et al. 2010; Turkheimer 2011). Yet, half of the heritability is still missing.

In quantitative genetics, heritability is defined as the portion of phenotypic variation in a population that is caused by genetic difference (Downes 2015). Traditionally, this portion is estimated by measuring the phenotypic resemblance of genetically related individuals without identifying at the molecular level (more particularly the DNA level) the genetic causes of phenotypic variation. GWAS have been developed in order to locate the DNA sequences that influence the target trait and estimate their effects, especially for common complex diseases such as obesity, diabetes and heart disease (Visscher et al. 2012; Frazer et al. 2009). As for height, almost 300 000 common DNA variants in human populations that associate with it have been identified by GWAS (Yang et al. 2010). Granted by many that the heritability estimates obtained

by traditional quantitative methods are quite reliable, the method(s) used in GWAS have been questioned (Eichler et al. 2010).

A number of partial solutions to the missing heritability problem have been proposed, with most of them focusing on improving the methodological aspects of GWAS in order to provide a more accurate estimate (e.g., Manolio et al. 2009; Eichler et al. 2010). Some authors have also suggested that heritable epigenetic factors might account for part of the missing heritability. For instance, in Eichler et al. (2000, 488), Kong notes that "[e]pigenetic effects beyond imprinting that are sequence-independent and that might be environmentally induced but can be transmitted for one or more generations could contribute to missing heritability." Furrow et al. (2011) also claim that "[e]pigenetic variation, inherited both directly and through shared environmental effects, may make a key contribution to the missing heritability." Others have made the same point (e.g., McCarthy and Hirschhorn 2008; Johannes et al. 2008). Yet, in the face of this idea one might notice what appears to be a contradiction: how can *epigenetic* factors account for the missing heritability, if the heritability is about *genes*?

To answer this question as well as to analyze the missing heritability problem, we compare the assumptions underlying both heritability estimates in traditional quantitative methods and those in GWAS. We argue that a) the heritability estimates of traditional methods include some terms associated with broad-sense heritability (H^2), as opposed to narrow-sense heritability (h^2); b) although GWAS are supposed to get h^2, h^2 relies on an evolutionary concept of the gene

that can include epigenetic factors while heritability estimates obtained from GWAS do not. With these two points being illustrated, we expect the missing heritability problem to be largely dissolved as well as setting the stage for further discussions.

The reminder of the paper will be divided into three parts. First, we briefly introduce how heritability is estimated in two traditional methods, namely twin studies and parent-offspring regression. We show that the estimates obtained by each methods include *some* non-additive elements and consequently correspond neither to H^2 nor to h^2, but to a notion in between which we term "broader-sense heritability". Second, we outline the basic rationale underlying GWAS and illustrate that they estimate heritability by considering solely DNA variants. By arguing that the notion of additive genetic variance does not necessarily refer to DNA sequences but can also refer to epigenetic factors in traditional quantitative methods, we show that the notion of heritability estimated in GWAS is more restrictive than that of traditional quantitative methods, and term this notion "DNA-based narrow-sense heritability". Finally, in Section 4, based on the conclusions from Section 2 and Section 3, we claim that the gap between the heritability estimates of traditional quantitative methods and those of GWAS can be explained away in two major ways. One consists in recognizing that if non-additive variance was removed from the estimates obtained via traditional methods, they would be lower. The other consists in recognizing that if epigenetic factors were taken into account by GWAS, the heritability estimates obtained would be higher. We conclude Section 4 by showing how our analysis sheds

some light on a discussion about the role played by non-additive factors in the missing heritability problem. Because human height has been "the poster child" of the missing heritability problem (Turkheimer 2011, 232), we will use this example to illustrate each of our points.

2. Heritability in Traditional Quantitative Methods.

According to quantitative genetics, the phenotypic variance (V_P) of a population can be explained by two components, its genotypic variance (V_G) and its environmental variance (V_E). In the absence of gene-environment interaction and correlation, we thus have:

$$V_P = V_G + V_E \quad (1)$$

From there broad-sense heritability (H^2) is defined as:

$$H^2 = \frac{V_G}{V_P} \quad (2)$$

V_G can further be portioned into the additive genetic variance (V_A), the dominance genetic variance (V_D) and the epistasis genetic variance (V_I). We have:

$$V_P = V_A + V_D + V_I + V_E \quad (3)$$

where V_A is the variance due to hypothetical genes making an equal and additive contribution to the trait studied (e.g., height). V_D is the variance due to interactions between alleles at one locus for diploid organisms, and V_I is the variance due to interactions between alleles from different loci. V_D and V_I together represent the variance due to particular combinations of genes of an organism.

Since genotypes of sexual organisms recombine at each generation via reproduction, dominance and epistasis effects are not transmitted stably across generations, only additive genetic effects are. Therefore, V_A is the variance due to stably transmitted genetic effects. Narrow-sense heritability (h^2) measures to what extent variation in phenotypes is determined by the variation in genes transmitted from parent(s) to offspring (Falconer and Mackay 1996, 123). It is defined as:

$$h^2 = \frac{V_A}{V_P} \quad (4)$$

h^2 is important in breeding studies and is used by evolutionary theorists who are interested in making evolutionary projections of a trait within a population across generations.

To know h^2, both V_A and V_P must be known. V_P, for most quantitative traits (including height), can be directly estimated by measuring individuals. However, there is no direct way to estimate V_A in traditional quantitative methods. The traditional way to estimate it requires two elements. First, one needs a population-level measure of a phenotypic resemblance of family

relative pairs[1]. This measure is obtained by calculating the *covariance* of the phenotypic values for those pairs. The choice of what sort of relatives to use depends on what data is available. The second element is the genetic relation between family pairs. It indicates the percentage of genetic materials the pairs are expected to share. With these two elements, one can estimate how much the genes shared contribute to the phenotypic resemblance. In a large population with different phenotypes, one can then estimate how much the additive genetic difference contributes to phenotypic difference in this population, which estimates h^2.

For simplicity, traditional quantitative methods usually assume that there is neither gene-environment interaction nor correlation (Falconer and Mackay 1996, 131). Thus the covariance between the phenotypic values (e.g., height) of pairs equals to additive genetic covariance, dominant and epistasis genetic covariance, plus the environmental covariance. A general equation for traditional quantitative methods can be written as follows:

$$Cov(P_1, P_2) = Cov(A_1 + D_1 + I_1 + E_1, A_2 + D_2 + I_2 + E_2) =$$
$$Cov(A_1, A_2) + Cov(D_1, D_2) + Cov(I_1, I_2) + Cov(E_1, E_2) \quad (5)$$

where indexes "1" and "2" represent the two family members for each pair studied. $Cov(P_1, P_2)$ is the covariance between the phenotypic values of one individual with the other.

[1] Or the mean values of their class (e.g., offspring) depending on the particular method used.

A, D, I and E represent additive effects, dominant effects, epistasis effects and environmental effects respectively.

The most commonly used traditional methods for estimating heritability are twin studies. In these studies one already knows that monozygotic twins share almost 100% of their genetic material while dizygotic twins about 50%. The environment is typically divided into the part of the environment that affects both twins in the same way (the shared environment, C) and the part of the environment that affects one twin but not the other (the unique environment, U) (Silventoinen et al. 2003). Hence, in the absence of interaction and correlation between C and U, we have:

$$E = C + U \quad (6)$$

Assuming epistasis effects to be negligible (a common assumption in twin studies), by inserting Equation (6) into Equation (5), we have:

$$Cov(P_{T1}, P_{T2}) = Cov(A_{T1} + D_{T1} + C_{T1} + U_{T1}, A_{T2} + D_{T2} + C_{T2} + U_{T2}) =$$
$$Cov(A_{T1}, A_{T2}) + Cov(D_{T1}, D_{T2}) + Cov(C_{T1}, C_{T2}) + Cov(U_{T1}, U_{T2}) \quad (7)$$

where indexes "T1" and "T2" represent the two twins for each twin pair studied. $Cov(P_{T1}, P_{T2})$ is the covariance between the phenotypic values of one twin with the other.

Because each twin's unique environment by definition is independent of that of the other twin, $Cov(U_{T1}, U_{T2})$ is zero for both monozygotic and dizygotic twins. Given that variance is a special case of covariance when the two variables are identical, and that for monozygotic twins A_{T1}, D_{T1}, and C_{T1} equal to A_{T2}, D_{T2}, and C_{T2} respectively, we can formulate the equation from Equation (7) as follows:

$$Cov_{MT}(P_{T1}, P_{T2}) = V_A + V_D + V_C \quad (8)$$

where $Cov_{MT}(P_{T1}, P_{T2})$ is the covariance between the phenotypic values of monozygotic twin pairs studied.

By contrast, dizygotic twins are expected to share half of their genes, which means that the covariance between the phenotypic values of one twin with the other of dizygotic twin pairs studied $(Cov_{DT}(P_{T1}, P_{T2}))$ is expected to be equal to half of the additive genetic variance, a quarter of dominant variance [2], and all of the shared environmental variance (with $Cov(U_{T1}, U_{T2})$ also to be zero). We have:

$$Cov_{DT}(P_{T1}, P_{T2}) = \frac{1}{2}V_A + \frac{1}{4}V_D + V_C \quad (9)$$

It is classically assumed that V_C in Equation (8) and (9) is the same. That is to say, for both monozygotic and dizygotic twin pairs, it is assumed that the shared environment would act in

[2] For each given gene with two alleles, the possibility that dizygotic twins have the same genotype is one quarter.

the same way if the pair has been reared together.[3] V_C can be cancelled by subtracting Equation (9) from Equation (8). The heritability can then be estimated as follows:

$$h^2_{b_{TS}} = \frac{2\{Cov_{MT}(P_{T1},P_{T2}) - Cov_{DT}(P_{T1},P_{T2})\}}{V_P} = \frac{V_A}{V_P} + \frac{\frac{3}{2}V_D}{V_P} \quad (10)$$

We call $h^2_{b_{TS}}$ broader-sense heritability (the index "b" is for "broader-sense") from *twin studies*, because the resulting estimate (which is about 0.8 for height) provides an accurate estimate of neither H^2 nor h^2, although it is closer to H^2 than to h^2 (Falconer and Mackay 1996, 172). That is to say, it corresponds to a definition of heritability that includes *some* elements of broad-sense heritability but not all of it.

Another often used traditional quantitative method to estimate heritability involves a parent-offspring regression. This method also assumes neither gene-environment interaction nor correlation, the covariance between the height of parents (one or the mean of both) and the mean of their offspring (Falconer and Mackay 1996, 164), equals to additive genetic covariance, dominant covariance (the epistasis covariance is assumed to be small and is not included), plus environmental covariance. Hence, Equation (5) can be formulated as follows:

[3] This assumption might be problematic because monozygotic twins are often treated more similarly by their parents than are dizygotic twins, and monozygotic twins are more likely to share a placenta than dizygotic twins. The difficulty can be mitigated by using adoption twin studies in which the environments for twins are random on average. But large adoption twins' data are exceedingly difficult to get (Griffiths 2005).

$$Cov(P_P, P_O) = Cov(A_P + D_P + I_P + E_P, A_O + D_O + I_O + E_O) =$$

$$Cov(A_P, A_O) + Cov(D_P, D_O) + Cov(E_P, E_O) \quad (11)$$

where indexes "P" and "O" represent the "parents" and the "offspring".

Two assumptions are then made. The first one is that there is no dominant effects transmitted from the parents to the offspring assuming the parents are unrelated (Doolittle 2012, 178), which means $Cov(D_P, D_O)$ is nil. Another assumption is that there is no correlation between the parents' environment and the offspring's environment so that $Cov(E_P, E_O)$ in Equation (11) is also nil. Given that on average, parents share in expectation 50% of genes with their offspring (parents and offspring share half of their genes), it leaves Equation (11) with a result of half of additive genetic variance ($\frac{1}{2}V_A$). Given V_P, h^2 can be estimated straightforwardly.

But the above two assumptions are problematic. First, the assumption of unrelated parents might be violated because of assortative mating in humans resulting in parents to be more genetically similar than two randomly chosen individuals (Guo et al. 2014). Hence, $Cov(D_P, D_O)$ is likely to be non-nil. Second, because the environments experienced by individuals are likely to be more similar within a family line, $Cov(E_P, E_O)$ might not be nil, either. If we take these two factors into consideration, the covariance of the parents and their

offspring is equal to half of additive genetic variance, *plus* a variance term representing effects due to dominance and similarities between environments. This can be written formally as:

$$Cov(P_P, P_O) = Cov(A_P, A_O) + Cov(D_P, D_O) + Cov(E_P, E_O) = \frac{1}{2}V_A + V_{D\&EC} \qquad (12)$$

where $V_{D\&EC}$ represents the variance due to some dominance and environmental correlation effects between the parents and the offspring studied.

The heritability can then be estimated by doubling the parent–offspring covariance in Equation (12) and dividing the total phenotypic variance of the population as follows:

$$h^2_{b_{POR}} = \frac{2Cov(P_P,P_O)}{V_P} = \frac{V_A}{V_P} + \frac{2V_{D\&EC}}{V_P} \qquad (13)$$

For similar reasons as with the heritability estimates from twin studies, we call $h^2_{b_{POR}}$ broader-sense heritability (with the index "b" also being for "broader-sense") from *parent-offspring regression*. Indeed, although it is often assumed that $h^2_{b_{POR}}$ represent h^2 (Falconer and Mackay 1996, 147), the resulting estimate (also about 0.8 for height) is broader than h^2 as it can include a component led by dominance variance and environmental correlation between parent and offspring.

To conclude this section, heritability estimates in both twin studies and parent-offspring regression include an extra term when compared to h^2, but they do not correspond to H^2. For this reason we regroup them under the term h^2_b for "broader-sense heritability", such that:

$$h_b^2 = h^2 + h_{other}^2 \quad (14)$$

where h_{other}^2 is the part of heritability contributed by the extra component(s) representing non-additive variance.

3. **Heritability in GWAS.**

Although any two unrelated individuals share about 99.5% of their DNA sequences, their genomes differ at specific nucleotide locations (Aguiar and Istrail 2013). Given two DNA fragments at the same locus of two individuals, if these fragments differ at a single nucleotide, they represent two variants of a Single Nucleotide Polymorphism (SNP). GWAS focus on SNPs across the whole genome that occur in the population with a probability larger than 1% which are called common SNPs. If one variant of a common SNP, compared to another one, is associated with a significant change on the trait studied, then this SNP is a marker for a DNA region (or a gene) that leads to phenotypic variation. For a polygenic trait like height, if we can detect all the SNPs that associate with it, then all the DNA difference makers that determine height difference can be located.

The development of commercial SNP chips makes it possible to rapidly detect common SNPs of DNA samples from all the participants involved in a study. By using a series of statistical tests, it can be investigated at the population level whether each SNP associates with

that target trait. The choice of the statistical tests depends on the data available as well as the trait studied. For quantitative traits like height, the most common approach is to make an analysis of variance table and assess whether the mean height of a group with one variant at one nucleotide is significantly different from the group with another variant of the same SNP[4] (Bush and Moore 2012). With all the SNPs associated with height being detected, data from the HapMap project, which provides a list of SNPs that are markers for most of the common DNA variants in human populations (Consortium, International HapMap 3 2010), is used to map the associated SNPs with common DNA variants. These mapped DNA variants, to be distinguished from DNA variants that do not affect the target trait, have been called "causal variants" (Visscher et al. 2012).

Based on the readings of SNP chips as well as further independent tests for SNPs, the effects of the associated SNPs (markers for causal DNA variants) on the trait can be calculated. By estimating the phenotypic variance contributed by these SNPs and the total phenotypic variance of the population, the heritability of causal DNA variants can be estimated as the ratio of the phenotypic variance caused by all the associated SNPs compared to the total phenotypic variance of the population (Weedon et al. 2008). Since it is common for biologists to assume

[4] For categorical (often binary disease/control) traits, the association test used involves measuring an odds ratio, namely the ratio of the odds of disease for individuals having a specific variant of a SNP, and the odds of disease for individuals who have another variant at the same locus. If the odds ratio of a common SNP is significantly different from 1, then that SNP is considered to be associated with the disease (Bush and Moore 2012).

that genes are only made up of pieces of DNA, it is thought that the variance obtained from all the causal DNA variants represent exactly the additive genetic variance, and the heritability estimated by GWAS should match narrow-sense heritability (h^2) (Yang et al. 2010; Visscher et al. 2006). However, the assumption that additive genetic effects are solely based on DNA sequences is problematic when faced with the evidence of epigenetic inheritance.

As was mentioned in Section 2, traditional quantitative methods for estimating heritability are based on measuring phenotypic values and genetic relations without reaching the molecular level. The genes are not defined physically, but functionally as heritable difference makers (Falconer and Mackay 1996, 123). In other words, they are theoretical units defined by their effects on the phenotype. With the discovery of DNA structure in 1953, it was thought that the originally theoretical genes were found in the physical DNA molecules. Since then, biologists commonly refer to genes as DNA molecules and this assumption is also made by researchers of GWAS. As [author] claim, this step was taken too hastily. If there is physical material, other than DNA pieces, that can affect the phenotype and be transmitted stably across generations, then it should also be thought to play the role that contributes to additive genetic effects.

Many studies have provided evidence for epigenetic inheritance[5], namely the stable transmission of epigenetic modifications across multiple generations and affect organism's traits

[5] We use the notion of "epigenetic inheritance" in the broad sense that refers to the inheritance of phenotypic features via causal pathways other than the inheritance of nuclear DNA (Griffiths and Stotz 2013, 112).

(e.g., Youngson and Whitelaw 2008; Dias and Ressler 2014). A classical example of this is the methylation pattern on the promoter of the agouti gene in mice (Morgan et al. 1999). It shows that mice with the same genotype but different methylation levels display a range of colors of their fur, and the patterns of DNA methylation can be inherited through generations causing heritable phenotypic variations. Epigenetic factors such as self-sustaining loops, chromatin modifications and three-dimensional structures in the cell can also be transmitted over multiple generations (Jablonka et al. 2014). Studies on various species suggest that epigenetic inheritance is likely to be 'ubiquitous' (Jablonka and Raz 2009).

The increasing evidence of epigenetic inheritance seriously challenges the restriction of the concept of the gene in the evolutionary sense to be materialized only in DNA. Relying on traditional quantitative methods, it is impossible to distinguish whether additive genetic variance is DNA based or based on other material(s). Some transmissible epigenetic factors, which are not DNA based, might *de facto* be included into the additive genetic variance used to estimate h^2. This extension of heritable units also echoes to the recent suggestion that genetic (assuming genes to be DNA based) and non-genetic heredity should be unified in an inclusive inheritance theory (Danchin 2013; Day and Bonduriansky 2010).

To apply the idea that some epigenetic factors can lead to additive genetic effects, the additive variance of them ($V_{A_{epi}}$) should be added to the additive variance of DNA sequences ($V_{A_{DNA}}$) to obtain V_A. Assuming there is no interaction between $V_{A_{epi}}$ and $V_{A_{DNA}}$, we have:

$$V_A = V_{A_{DNA}} + V_{A_{epi}} \quad (15)$$

Inserting Equation (15) to Equation (4) leads to:

$$h^2 = \frac{V_{A_{DNA}}}{V_P} + \frac{V_{A_{epi}}}{V_P} \quad (16)$$

Here we term the first term on the right side of Equation (16) "DNA-based narrow-sense heritability" (h^2_{DNA}), and the second term "epigenetic-based narrow-sense heritability" (h^2_{epi}), we thus have:

$$h^2_{DNA} = h^2 - h^2_{epi} \quad (17)$$

4. **Dissolving the Missing Heritability**.

As we mentioned it in Introduction, since the first successful GWAS was published in 2005 (Klein et al. 2005), there have been a lot of proposals for methodological improvements in GWAS (Manolio et al. 2009; Eichler et al. 2010). Studies have been conducted according to those proposals that permit to obtain higher heritability estimates. Examples include increasing the sample sizes which has resulted in more accurate estimates (e.g., Wood et al. 2014), considering all common SNPs simultaneously instead of one by one which has increased the heritability estimates of height from 0.05 to 0.45 (see Yang et al. 2010), and conducting meta-analyses which can lead to more accurate results when compared to single analysis (see Bush

and Moore 2012). Biologists have also suggested to search for SNPs with lower frequencies than 1% in order to account for a wider range of possible causal variants (Schork et al. 2009).

Aside from these partial improvements, our analysis reveals two reasons explaining away the missing heritability problem: a) In traditional quantitative methods, the heritability estimates include extra terms which are not presented in GWAS; b) In GWAS, heritability is estimated solely from causal DNA variants, while in traditional quantitative methods the additive effects contributed by epigenetic difference (h^2_{epi}) are *de facto* included in the estimates.

These two reasons can be shown formally. Using our terminology, missing heritability (MH) equals to the estimates obtained by traditional quantitative methods (h^2_b) minus the estimates obtained by GWAS (h^2_{DNA}), which are 0.8 and 0.45 respectively in the case of height. Thus we have:

$$MH = h^2_b - h^2_{DNA} \tag{18}$$

Replacing h^2_b and h^2_{DNA} by the right hand side of Equation (14) and (17), we obtain:

$$MH = h^2_b - h^2_{DNA} = h^2 + h^2_{other} - \left(h^2 - h^2_{epi}\right) = h^2_{other} + h^2_{epi} \tag{19}$$

Which means that the missing heritability results from the part of heritability originating from epigenetic factors stably transmitted across generations, plus the part of heritability originating from non-additives factors.

Our point that part of the missing heritability can be dissolved by considering non-additive effects echoes to the claim that almost all GWAS to date have focused on additive effects might be a reason for the missing heritability (McCarthy and Hirschhorn 2008). Although there is not enough data to confirm that non-additive effects do explain away some part of missing heritability, this claim appears again and again in discussions on the missing heritability problem (see for instance Maher 2008; Frazer et al. 2009; Gibson 2010; Kong 2010; Moore 2010). Yang et al. (2010, 565) disagree with this claim and respond that "[n]on-additive genetic effects do not contribute to the narrow-sense heritability, so explanations based on non-additive effects are not relevant to the problem of missing heritability."

We agree with Yang et al. (2010) that non-additive effects do not contribute to h^2. That said, because the heritability estimates obtained from traditional quantitative methods do not strictly correspond to h^2 but include some non-additive elements, non-additive effects cannot be dismissed as irrelevant for the missing heritability problem, though probably they are relevant in a way that both Yang et al. (2010) as well as their opponents did not consider.

5. **Conclusion**.

We have provided two ways in which the missing heritability problem can be explained away. First, heritability estimates from traditional quantitative methods (h_b^2) are overestimated when

compared to h^2. The resulting estimates would be smaller if the non-additive elements were eliminated. Second, heritability estimates from GWAS (h^2_{DNA}) are underestimated when compared to h^2 because they do not take into account the additive effects of epigenetic factors behaving like evolutionary genes. The resulting estimates would be larger if epigenetic factors were taken into account. We have voluntarily stayed away from the question of whether heritability should be defined strictly relative to DNA sequences or if it should encompass any factors behaving effectively like an evolutionary gene. Our inclination is that there is no principled reason to exclude non-DNA transmissible factors from heritability measures, but our analysis does not bear on this choice.

References:

Aguiar, Derek, and Sorin Istrail. 2013. "Haplotype Assembly in Polyploid Genomes and Identical by Descent Shared Tracts." *Bioinformatics* 29 (13): i352–i360.

Authors. Forthcoming. "The Evolutionary Gene and the Extended Evolutionary Synthesis." *British Journal for Philosophy of Science*.

Bush, William S., and Jason H. Moore. 2012. "Genome-Wide Association Studies." *PLoS Computational Biology* 8 (12): e1002822.

Consortium, International HapMap 3. 2010. "Integrating Common and Rare Genetic Variation in Diverse Human Populations." *Nature* 467 (7311): 52–58.

Danchin, Étienne. 2013. "Avatars of Information: Towards an Inclusive Evolutionary Synthesis." *Trends in Ecology & Evolution* 28 (6): 351–358.

Day, Troy, and Russell Bonduriansky. 2011. "A Unified Approach to the Evolutionary Consequences of Genetic and Nongenetic Inheritance." *The American Naturalist* 178 (2): E18–E36.

Dias, Brian G., and Kerry J. Ressler. 2014. "Parental Olfactory Experience Influences Behavior and Neural Structure in Subsequent Generations." *Nature Neuroscience* 17 (1): 89–96.

Doolittle, Donald P. 2012. *Population Genetics: Basic Principles*. Vol. 16. Springer Science & Business Media.

Downes, Stephen M. 2015. "Heritability." In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University.

Eichler, Evan E., Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore, and Joseph H. Nadeau. 2010. "Missing Heritability and Strategies for Finding the Underlying Causes of Complex Disease." *Nature Reviews Genetics* 11 (6): 446–450.

Falconer, Douglas S., and Trudy F. C. Mackay. 1996. *Introduction to Quantitative Genetics*. 4th edition. Longman: Benjamin Cummings.

Feil, Robert, and Mario F. Fraga. 2012. "Epigenetics and the Environment: Emerging Patterns and Implications." *Nature Reviews Genetics* 13 (2): 97–109.

Frazer, Kelly A., Sarah S. Murray, Nicholas J. Schork, and Eric J. Topol. 2009. "Human Genetic Variation and Its Contribution to Complex Traits." *Nature Reviews Genetics* 10 (4): 241–251.

Furrow, Robert E., Freddy B. Christiansen, and Marcus W. Feldman. 2011. "Environment-Sensitive Epigenetics and the Heritability of Complex Diseases." *Genetics* 189 (4): 1377–1387.

Griffiths, Anthony JF., Susan R. Wessler, Richard C. Lewontin, William M. Gelbart, David T. Suzuki, and Jeffrey H. Miller. 2005. *An Introduction to Genetic Analysis*. 8th edition. New York: W. H. Freeman.

Griffiths, Paul, and Karola Stotz. 2013. *Genetics and Philosophy: An Introduction*. Cambridge University Press.

Guo, Guang, Lin Wang, Hexuan Liu, and Thomas Randall. 2014. "Genomic Assortative Mating in Marriages in the United States." *PloS One* 9 (11): e112322.

Jablonka, Eva, Marion J Lamb, and Anna Zeligowski. 2014. *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Revised edition. MIT Press.

Jablonka, Eva, and Gal Raz. 2009. "Transgenerational Epigenetic Inheritance: Prevalence, Mechanisms, and Implications for the Study of Heredity and Evolution." *The Quarterly Review of Biology* 84 (2): 131–176.

Johannes, Frank, Vincent Colot, and Ritsert C. Jansen. 2008. "Epigenome Dynamics: A Quantitative Genetics Perspective." *Nature Reviews Genetics* 9 (11): 883–890.

Klein, Robert J., Caroline Zeiss, Emily Y. Chew, Jen-Yue Tsai, Richard S. Sackler, Chad Haynes, Alice K. Henning, John Paul SanGiovanni, Shrikant M. Mane, and Susan T. Mayne. 2005. "Complement Factor H Polymorphism in Age-Related Macular Degeneration." *Science* 308 (5720): 385–389.

Maher, Brendan. 2008. "Personal genomes: The Case of the Missing Heritability." *Nature News* 456 (7218):18–21.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, and Aravinda Chakravarti. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461 (7265): 747–753.

McCarthy, Mark I., and Joel N. Hirschhorn. 2008. "Genome-Wide Association Studies: Potential next Steps on a Genetic Journey." *Human Molecular Genetics* 17 (R2): R156–165.

Morgan, Hugh D., Heidi GE Sutherland, David IK Martin, and Emma Whitelaw. 1999. "Epigenetic Inheritance at the Agouti Locus in the Mouse." *Nature Genetics* 23 (3): 314–318.

Schork, Nicholas J., Sarah S. Murray, Kelly A. Frazer, and Eric J. Topol. 2009. "Common vs. Rare Allele Hypotheses for Complex Diseases." *Current Opinion in Genetics & Development* 19 (3): 212–219.

Silventoinen, Karri, Sampo Sammalisto, Markus Perola, Dorret I. Boomsma, Belinda K. Cornes, Chayna Davis, Leo Dunkel, Marlies De Lange, Jennifer R. Harris, and Jacob VB

Hjelmborg.2003. "Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries." *Twin Research* 6 (05): 399–408.

Turkheimer, Eric. 2011. "Still Missing." *Research in Human Development* 8 (3-4): 227—241.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. "Five Years of GWAS Discovery." *The American Journal of Human Genetics* 90 (1): 7–24.

Visscher, Peter M., Sarah E. Medland, Manuel AR Ferreira, Katherine I. Morley, Gu Zhu, Belinda K. Cornes, Grant W. Montgomery, and Nicholas G. Martin. 2006. "Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings." *PLoS Genet* 2 (3): e41.

Visscher, Peter M., William G. Hill, and Naomi R. Wray. 2008. "Heritability in the Genomics Era—concepts and Misconceptions." *Nature Reviews Genetics* 9 (4): 255–266.

Weedon, Michael N., Hana Lango, Cecilia M. Lindgren, Chris Wallace, David M. Evans, Massimo Mangino, Rachel M. Freathy, John RB Perry, Suzanne Stevens, and Alistair S. Hall. 2008. "Genome-Wide Association Analysis Identifies 20 Loci that Influence Adult Height." *Nature Genetics* 40 (5): 575–583.

Wood, Andrew R, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, and Zoltán Kutalik. 2014. "Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height." *Nature genetics* 46 (11): 1173–1186.

Yang, Jian, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, et al. 2010. "Common SNPs Explain a Large Proportion of the Heritability for Human Height." *Nature Genetics* 42 (7): 565–569.

Youngson, Neil A., and Emma Whitelaw. 2008. "Transgenerational Epigenetic Effects." *Annual Review of Genomics and Human Genetics* 9: 233–257.

Scientific expertise, risk assessment, and majority voting

Thomas Boyer-Kassem*

Working paper: comments welcome, but please do not quote without permission.

February 29, 2016

Abstract

Scientists are often asked to advise political institutions on pressing risk-related questions, like climate change or the authorization of medical drugs. Given that deliberation will often not eliminate all disagreements between scientists, how should their risk assessments be aggregated? I argue that this problem is distinct from two familiar and well-studied problems in the literature: judgment aggregation and probability aggregation. I introduce a novel decision-theoretic model where risk assessments are compared with acceptability thresholds. Majority voting is then defended by means of robustness considerations.

Keywords: scientific expertise, risk, majority voting, robustness, decision theory

*TiLPS, Tilburg University, The Netherlands. Email: t.c.e.boyer-kassem@uvt.nl

1 Introduction

Scientists are often asked by political institutions to give expert advice on pressing questions. For instance, agencies that regulate medicines regularly resort to expert panels, and national scientific academies give advice to the government or to the assemblies. Even after discussing, scientific experts do not always agree on the answer, and when they do, they may disagree on the justification for this answer. How should decisions that involve risk assessments be taken and justified within scientific expert panels? This is the central question studied in this paper. As a matter of fact, many expert panels take decisions using the majority voting rule. This is for instance the case in advisory committees in the European and in the American agencies that grant medicines authorization, respectively the EMA and the FDA.[1] But is it the best decision rule? Is majority voting *on the final decision* the best way to aggregate different experts' opinions, and to track their reasons? This paper is restricted to cases in which the expert panel is asked to take a decision on only one binary question, for instance to answer the question "Is the risk-benefit ratio of some medicine worth it to be authorized for commercial use?". This simple case is already interesting as it corresponds to many real-life cases: some expert panels are constituted on the sole purpose of answering one specific question, or are asked to answer several but logically unrelated questions — e.g. decisions about different medicines.

To study this problem, I introduce a novel decision-theoretic model. The true/false decision is supposed to be taken by comparing a risk assessment a (typically, a probability) to a risk acceptability threshold t, e.g. "true" if and only if $a < t$. For simplicity, a and t are supposed to be in $[0, 1]$, but any quantity might go.[2] It is assumed that the n experts agree on the threshold value, but differ in their individual risk assessments a_k $(k = 1, \ldots, n)$ — or conversely, that they agree on the assessment, but disagree on the threshold value. Typically, the question asked to the expert panel is in the form "Is X's risk below t?". The problem studied in this paper is to determine how the individual a_k's should be aggregated in comparison with t, so as to give the group's answer to this question (I shall speak equivalently of the group's decision, or of the group's belief on whether the risk is below t). Compared to probability aggregation theory which studies the aggregation of probabilistic opinions, the novelty of this model lies (i) in the introduction of a threshold comparison which projects probabilities into a binary space, and (ii) in the fact that the group has to take a

[1] Cf. Hauray and Urfalino (2007), Urfalino and Costa (2015).

[2] Real quantities can be mapped to the interval $[0, 1]$, for instance with the function $x \to 1 - 1/(1 + x)$).

stand on one binary question only, and not on a more complex agenda. Compared to judgment aggregation theory which studies the aggregation of an interconnected set of beliefs, the novelty is that individuals do not just have true/false beliefs but probabilistic ones, even if the group is asked to express a true/false belief in the end. The present problem can be considered as a first bridge between these two existing frameworks. The best decision rule for our binary question is likely to depend on the details of characteristics of the question, of the experts, of the available knowledge, and on other details. My methodological approach is not to conduct a detailed study of particular cases, but to look at features which most (interesting) cases share, so as to find general properties of the best decision rule — what is meant by "best" shall be discussed too.

The main claims of this paper are the following. I argue that the framework of probability aggregation cannot help us solve the present problem (Section 2), because the aggregation problems it considers are too general. For the aggregation of scientific risk assessment on a specific question, a theory of its own is needed, and I try to sketch one here. I then argue that robustness considerations clearly legitimate majority voting on the final decision (Section 3). But when justifications for the decisions are sought, majority voting can lead to inconsistencies and the expert panel should aggregate on the reasons separately, before deriving logically its decision (Section 4). Overall, the case for the majority rule is thus a mixed one.

2 Probability Aggregation and Beyond

A standard requirement for a scientific expert panel is that it provides justifications for its decision. In the present model, the decision has to be consistent with the comparison between the risk assessment and the threshold, so a minimal justification is that the panel has a belief on the risk assessment (as all experts have a belief on the risk assessment, it would be weird that the panel claims to refuse the authorization while not being able to say that it believes that the risk assessment is above the threshold). So, our problem includes as a first step the aggregation of the individual risk assessments $\{a_k\}_{1\leq k\leq n}$ into a single group assessment a — deeper justifications for the group's decision are contemplated in Section 4. The group's decision is supposed to be consistent with this assessment, so pragmatically the easiest way to do so may be for the group to first aggregate the individual assessments, and then compare the result to the threshold.

Majority voting on the decision itself is a standard way for expert groups to take decisions, but it does not proceed in that way. Can it be objected that, within our model, it lacks the requirement that the group should be attributed a belief

on the risk assessment? No, for the following reason. The result of the majority vote is "true" if and only if a majority of agents vote "true", i.e. if and only if a majority of agents have a numerical assessment below the threshold, i.e. if and only if the median of the agents' assessments is below the threshold. In other words, the majority voting rule on the decision is equivalent to considering that the group's assessment is the median of the individual assessments. Hence, majority voting is in the race. What are the other challengers? A standard way to aggregate probabilities is to make averages. The linear average is defined as $\sum_k a_k$, and it can be generalized with weights $\omega_k \geq 0$ and $\sum_k \omega_k = 1$, as $\sum_k \omega_k a_k$, to take into account unequal degrees of expertise on the question.[3] Other averages are the geometric average or the harmonic average. Our problem is to determine which probability aggregation rule, followed by the threshold, is the best one in our problem. It is easy to see that these various probability aggregation rules can give different binary decisions for the group.[4]

Pooling probability functions has been studied for several years in the theory of probability aggregation (for surveys, cf. Dietrich and List forth., Martini and Sprenger forth., section 3). Can its results be used to select the best aggregation rule in our problem? I shall argue that unfortunately no. The framework of probability aggregation adopts an axiomatic method: it starts by stating several axioms which appear as desirable properties for the pooling function and then studies which function or aggregation rule, if any, satisfies them. The axioms considered in Dietrich and List's survey can be expressed in our case as:

- **Independence**: the group's probability a only depends on the individual probabilities a_k.
- **Unanimity preservation**: if all agents' probabilities a_k are the same, then the group's probability a is this one too.
- Three **Bayesian axioms**: if some information is learned by all individuals, then the group's decision changes by conditionalization on that event.

[3] It is akin to the iterated Lehrer-Wagner model which, starting from respect weights agent have to one another, provides a single probability for the group. However, the iterated Lehrer-Wagner model, and even more its normative interpretation, have been subjected to many criticisms (for a survey, cf. e.g. Martini and Sprenger forth. section 4). As a descriptive model, it is not useful for the present discussion.

[4] Consider for instance the median and the linear average, with three experts with $a_1 = a_2 = 0.04$, $a_3 = 0.10$, and $t = 0.05$. A majority voting on the decision gives a "true" as two experts on three assess the risk to be below the threshold. The linear average (with equal weights) is 0.06, which is higher than t, so this gives a "false".

The Independence axiom is automatically satisfied here, because our problem contains only one true/false answer, and there is no other probability on which a could depend. The three Bayesian axioms make sense in cases where the expert panel learns new information. In our problem, however, an extensive discussion has already taken place so no agent learns new information anymore, and the expert panel is not making any new inquiry. So the Bayesian axioms are not relevant in our case, and only the Unanimity preservation axiom expresses a desirable property for the aggregation rule.

An essential point to note is that a very large number of aggregation rules satisfy this axiom: the median, linear averaging, geometric averaging, and so on — actually, any convex function of the a_k. This illustrates the fact that a classical uniqueness result from the probability aggregation literature does not hold anymore: the well-known theorem by McConway 1981 and Wagner 1982, which states that linear averaging functions are the *only* independent and unanimity-preserving functions. The reason is that the theorem requires a set of at least three events, whereas our problem only considers two — e.g. the product is risky, with probability a_k, and the product is not risk, with probability $1 - a_k$. Considering a simpler agenda has widened the set of suitable aggregation rules, and no theoretical result from the literature can be used to pick the best one. More generally, the uniqueness and impossibility results from the theory of probability aggregation are useless for our problem. So, how scientific expert panels should aggregate risk assessments is not a simple problem that can be solved straightforwardly with the existing literature, which has focused on general problems with complex agendas, and has thus neglected more specific yet important questions. In the next section, I discuss other desiderata or axioms that we would like to impose on the aggregation rule.

3 Robustness Matters

Scientific risk assessment is supposed to meet some standards of reliability and objectivity, and the aggregation of these assessments should follow alike standards. In this spirit, I now introduce several new requirements for our aggregation rule. The aggregation rule should be sensitive to the right features of our problem, and not to the parasitic ones. It should favor objective features at the detriment of idiosyncrasies or unwanted values (for an analysis of the concept of objectivity, cf. Douglas 2004 — I refer to some of her distinctions below). In other words, the aggregation rule should be robust to some changes that we regard as irrelevant. In this section, I defend three dimensions of robustness that should be taken into account: the risk metrics, the level of detail, and the presence of strategical agents.

Several probability aggregation rules can be considered: linear averaging, geometric averaging, harmonic averaging, among others. As the forthcoming robustness discussion is similar for all the various averagings, I shall simplify it and consider only linear averaging, which shall be contrasted with the median. $\mathcal{R}_a$ denotes the aggregation rule that compares the threshold with the linear average (which thus stands for other averages), and $\mathcal{R}_m$ the aggregation rule that compares the threshold with the median of the individual assessments (which is equivalent to a majority vote on the decision itself).

3.1 Metrics

The formal model I have introduced relies on a quantitative scale — a and t are given numerical values in $[0, 1]$. How is this scale defined in real cases? My talking about probabilities has been only a matter of simplicity given the reduction of the problem to the $[0, 1]$ interval, and typical cases do not bear on well-defined probabilities or explicit scales. For instance, a standard question posed at an FDA advisory committee is "Does the overall risk versus benefit profile for X support marketing in the US ?"[5]. This question supposes that experts identify the risk versus benefit profile, and determine the value of the threshold under which a marketing is warranted. This can be done in a number of ways, and these are essentially value-laden questions[6] — what is acceptable or not has to do with extra-scientific values, and may also reflect the fact that an expert is risk-averse or risk-seeking. Overall, it makes sense to suppose that both the metrics scale and the threshold depend on the experts. Conversely, as the aggregation procedure is supposed to take place when the experts have extensively discussed, one can make the simplifying assumption that the same facts are known to all, and thus that the risk assessment is the same for all. In that way, our model actually applies in the setting in which a is common to all experts, but each has her own threshold t_k. The fact that the quantitative risk scale is not uniquely defined can be approached from a mathematical viewpoint: any scale can be reparametrized by applying any continuous bijection from $[0, 1]$ to $[0, 1]$, such as $x \mapsto x^2$.

These points make a hard time for the rule $\mathcal{R}_a$ (and other non linear averagings). First, from a practical viewpoint, the dependence of the risk scale metrics on the expert prevents the use of rules which take as inputs the numerical values of the risk assessments or of the threshold. For instance, is it even possible for a chairman to ask her colleagues "Please tell me your overall risk versus benefit acceptability threshold"

[5] Cf. Urfalino and Costa (2015, p.183).

[6] On the role of values in science more generally, and a critic of the value-free ideal, cf. Douglas 2009.

(or assessment), given that each expert may have her own scale? The rule $\mathcal{R}_m$, as it is equivalent to majority voting, needs not rely on input individual numerical values, and is thus safe from this criticism. Second, even if these practical difficulties could be overcome, some theoretical difficulties remain. Suppose a common scale has been adopted so that all experts can express their t_k. An aggregation rule that depends on the metrics of that common scale can give different outcomes according to the scale employed, as shown in Table 1. This dependence is a problem: which common scale should be chosen? (This is another aggregation problem!) Note that a variant of this problem exists even with a well-defined probability scale. For instance, let A be the event that a certain risk (e.g. carcinogenic substances in food) is responsible for more than 10 cases of cancer in 100,000 people during 1 year. The experts estimate the probability of A, $p(A)$. Consider now A' the event that the risk is responsible for more than 10 cases of cancer in 100,000 people during 10 years. Call $p(A')$ its probability. If the cancer cases are independent along the years, then $p(A') = 1 - (1 - p(A))^{10}$. Because the relation between $p(A)$ and $p(A')$ is not linear, taking the linear average of the experts assessments on A, and transforming it into an assessment on A', or taking the linear average of the experts assessments on A', does not give the same result. Which event A or A' is the more "natural" is not clear, and so much more for the right risk group assessment.

This gives good reasons to consider the following requirement: the aggregation rule should be insensitive to the metrics used to describe the problem, i.e. the assessment and the threshold. What should matter is just the relative position of the a and t_k, not their distance which can be due to some idiosyncratic value-laden judgments. This is requiring that the aggregation rule is more objective, under the sense of value-neutral objectivity as characterized by Douglas (2004, p. 460), which does not mean "free from all value influence" (as judging whether a risk benefit ratio is lower enough is bound to involve a value judgment), but takes a position "that is balanced or neutral with respect to a spectrum of values" (here, the balance is reached by taking into account only relative positions). The metrics robustness excludes the rule $\mathcal{R}_a$ which employs a linear average — Table 1 has shown a counter-

	t_1	t_2	t_3	a	Average t	$\mathcal{R}_a$	$\mathcal{R}_m$
x scale	.01	.01	.1	.05	.04	**False**	False
x^2 scale	0.0001	0.0001	.01	0.0025	0.0034	**True**	False

Table 1: Example in which the rule $\mathcal{R}_a$ gives different answers depending on the scale. The three experts have different thresholds t_k and a common risk assessment a.

example — but not $\mathcal{R}_m$ which relies on the median.[7]

3.2 Level of detail

Another argument for an aggregation rule that does not rely on a specific metrics comes from considerations of the level of detail in which the problem is described. So far, a continuous scale has been assumed, with numerical assessments in $[0, 1]$. Numerical discrete scales could also be used or even qualitative assessments only — it corresponds to decisions under uncertainty and not under risk. Consider for instance the case of the well-known IPCC Assessment Reports, that formulate a synthesis of existing scientific knowledge on climate change issues. The reports use a standardized vocabulary to express uncertainties, with several scales: some are qualitative (e.g. low/medium/high), others are quantitative (and use probabilities).[8] The historical trend has been to use more quantitative scales and less qualitative scales, but the latter have the advantage of being easily understandable by non-technical audiences, and thus should continue to be used in the future. Some qualitative and quantitative scales are in an explicit correspondence, as illustrated on Table 2. Writing an IPCC report involves synthesizing large amounts of scientific literature, so co-authors of a chapter may have different beliefs on the uncertainties associated with a finding. Whether they express their beliefs on a qualitative or on a quantitative scale, the way their beliefs are aggregated should be smooth and not vary abruptly (some very precise yet qualitative scales are conceivable), all the more than some explicit correspondence exist (Table 2). This is also a question of historically

[7]The comparability of scales is also discussed in Risse's (2004) political philosophy work, who also takes it as an argument for majority voting.

[8]Cf. e.g. the last report of the Working Group I, Stocker et al (2013, p. 138-142).

Term	Likelihood of the Outcome
Virtually certain	99–100 % probability
Very likely	90–100% probability
Likely	66–100% probability
About as likely as not	33–66% probability
Unlikely	0–33% probability
Very unlikely	0–10% probability
Exceptionally unlikely	0–1% probability

Table 2: Likelihood terms associated with outcomes used in the Fifth Assessment Report of the IPCC (Stocker et al 2013, p. 142).

consistency when switching from qualitative to quantitative scales.[9] Thus, a sound requirement is that the aggregation rule extends to formulations with discrete and qualitative scales. As the average of non-numerical and qualitative values is not defined, $\mathcal{R}_a$ does not satisfy this requirement. The median is defined on any kind of scale, and $\mathcal{R}_m$ satisfies the requirement. So only $\mathcal{R}_m$ is robust for the level of detail.

3.3 Bias and strategical votes

Not all experts are moved by epistemic goals only, and conflicts of interests can arise. For instance, numerous controversies have surrounded the FDA advisory committees along the years (Urfalino and Costa 2015, p. 168-169.) If a better selection of experts may be the solution, the decision rule used in the expert panel can also reduce the impact of bias agents.[10] With $\mathcal{R}_a$, an expert can strategically express a much lower risk of a medicine to influence the group's average — with a threshold at 10 %, she might express 0.1% instead of just 9%. The aggregation rule should be insensitive to such a strategical vote manipulation, and this is all the more important as the biased agent may have already influenced other agents during the preceding discussion. $\mathcal{R}_m$ is clearly robust in this sense, as an agent has the same influence whether her probability is just below the threshold or close to 0. This is not so for $\mathcal{R}_a$. This robustness requirement also makes the aggregation rule more objective, in the sense of detached objectivity (Douglas 2004, p. 459): one's personal values (allegiance to a firm) should not prevail on evidence (e.g. that the probability is 9%, as above).

Overall, the three robustness requirements considered here clearly favor $\mathcal{R}_m$ over $\mathcal{R}_a$. This provides a substantial justification for the traditional democratic rule in expert panels confronted with a binary decision. This result is a real departure from probability aggregation theory, in which linear averaging is justified on solid grounds. Narrowing the agenda and introducing a threshold has changed the solution to the aggregation problem.

[9] One may object that in the IPCC case the co-authors aggregate beliefs without a threshold comparison for a binary decision. Actually, thresholds are implicit: a finding which confidence is too low may not be mentioned. Anyway, the IPCC example can be seen as a mere illustration of the level of detail problem.

[10] Biased and extremist agents have been much studied in the literature of opinion dynamics (cf. for instance in Lorenz's 2007 survey), but not so in the literature of opinion aggregation.

4 Reasons

So far, a simplified model of scientific expert panels has been considered, one in which the group is asked to give a binary decision. As argued, the first step in justifying that decision consists for the panel to have a belief on the risk assessment, which is given by the median of the individual assessments in the case of $\mathcal{R}_m$. However, expert panels are often asked to provide a deeper justification. The question then arises of how the panel should aggregate its members views on this justification. In this section, I propose a novel but simple model for individual numerical assessment justification, in line with my previous threshold model.

Perhaps the most typical interpretation of the risk assessment a is that of a (subjective) probability. Suppose this probability is determined by m independent factors ($m \geq 2$). For instance, the risk associated with a medicine comes from m unrelated secondary effects. Then a is the probability that at least one risk factor triggers:

$$a = 1 - \prod_{j=1}^{m} (1 - a_j). \tag{1}$$

Each expert k is supposed to have her own assessment of each factor $a_{k,j}$ ($j = 1, \ldots, m$). Our problem is then to aggregate the $n \times m$ matrix of probabilities $a_{k,j}$, and to compare that result with the threshold.

As the m factors are independent, a sound requirement is to aggregate the individual assessments on them separately. How should that be done? Adapting the arguments from the previous section, one is lead to the conclusion that the panel should take the *median* of the individual assessments for each factor. However, there is a fundamental limitation to this, due to the previously mentioned theorem by McConway and Wagner's (cf. Section 2). Here is why. Requiring as above that the aggregation proceeds on each factor independently is just requiring the classical independence axiom. Another legitimate requirement is the classical axiom of unanimity preservation: if all experts agree on the risk assessment for one factor, then the panel should take this value as its own. As $m \geq 2$, all the conditions of the theorem by McConway and Wagner are fulfilled[11], so its conclusion apply: the only probability aggregation rule on the set of factors and on the overall decision is linear averaging. This reveals that, if groups use the median to determine both the independence factors' values and the overall risk (according to the above results), then it does not give a probability function and inconsistencies can arise. Table 3

[11]Each of the $m \geq 2$ factors can be triggered or not, so there are at least 4 events, which is higher than the 3 required in the theorem.

gives such an example. In other words, asking the expert panel to take stands on the reasons for its majority decision can lead it to change its decision.

Does it mean that our robustness defense of the median should be discarded? Not necessarily. The theorem by McConway and Wagner assumes that the experts aggregate their views *both* on the independent factors and on the overall risk assessment. But one can have the experts aggregate their views on the independent factors only. The overall risk assessment is then computed according to Equation 1, and the final decision is logically obtained from a comparison between this value and the threshold. In that way, experts do not vote on the final decision directly. This decision rule is a so-called premise-based rule.[12] Then, the linearity result of McConway and Wagner does not apply any more. The robustness considerations from the previous section do apply at the level of independent factors, and they recommend that the group takes the median of the individual assessments.

The present model of factors has assumed that there exists some common numerical scale, so that taking the median of individual assessments makes sense. However, the previous section has in part argued that such a scale may not always exist. In these cases, the present model of independent factors cannot apply. The theory of judgment aggregation offers a general framework for the aggregation of non-numerical reasons or justifications, with true/false beliefs (for reviews, cf. List 2012, Martini and Sprenger forth.). Applying in detail this framework to our problem of scientific justification would require another paper. A general result from this literature, however, is the discursive dilemma: majority voting on a set of true/false beliefs related in a logical way (here: reasons for the decision) may generate inconsistent collective judgments. This echoes our own finding about the median, which corresponds to majority voting in case of a threshold comparison. So whatever

[12] On this strategy more generally, see Cooke (1991), Bovens and Rabinowicz (2006), Hartmann and Sprenger (2012). Another solution to our problem could be the conclusion-based rule, i.e. aggregate only the views on the conclusion, but this is just like the previous section that we are trying to surpass.

Risk aspect	a_1	a_2	$a = 1 - (1 - a_1) \cdot (1 - a_2)$
Agent #1	0.01	0.01	0.0199
Agent #2	0.02	0.01	0.0298
Agent #3	0.01	0.02	0.0298
Median	0.01	0.01	0.0199 or 0.0298 ?

Table 3: A case in which the rule of the median can lead to inconsistencies. With a threshold at e.g. 0.025, the group's decision could be either true or false.

the scale, majority voting on all parts of the question is in great difficulty, and a premise-based solution should be adopted.

5 Conclusion

This paper has investigated the rationale for the majority rule that is often used in scientific expert panels, when dissent persists after discussion, and has looked for the best decision rule in this context. To this end, I have introduced a threshold probability model for individual decisions. Three main points have been shown in the paper: (1) the standard framework of probability aggregation is unable to solve our problem of risk aggregation. (2) robustness considerations clearly favor majority voting on the decision, i.e. comparing the threshold to the median of the individual risk assessments. (The robustness axioms I have advocated, which have been designed from considerations on scientific expert panel, could in return inspire social choice theory). (3) when a justification of the panel's decision is looked for, the median rule (corresponding to majority voting) can lead to inconsistencies. The promising route is to have the group aggregate on the reasons level, not on the final decision one. This should encourage scientific expert panels to divide questions from a logical viewpoints, and to take decisions on sub-problems instead of voting on the final decision directly. Current practices in advisory committees of the FDA and of the EMA could evolve in this respect. However, these claims have only been shown in quite simple and idealized models of decision-making. Future work is needed to investigate other models. These preliminary results have nonetheless cast some serious doubts on the majority voting rule only applied on the final decision.

Note finally the generality of the proposed model, which goes well beyond scientific expertise: the a and t variables can be interpreted as degrees of beliefs or as utility measures, within an epistemology or an economy framework.

References

Bovens, Luc and Wlodek Rabinowicz. 2006. "Democratic answers to complex questions. An epistemic perspective". *Synthese* 150: 131-153.

Cooke, Roger M. 1991. *Experts in Uncertainty. Opinion and Subjective Probability in Science*. Oxford University Press.

Dietrich, Franz and Christian List. Forthcoming. "Probabilistic Opinion Pooling". In *Oxford Handbook of Probability and Philosophy*, Oxford University Press.

Douglas, Heather E. 2004. "The Irreducible Complexity of Objectivity." *Synthese* 138:453-473.

Douglas, Heather E. 2009. *Science Policy and the Value-Free Ideal*. University of Pittsburgh Press.

Hartmann, Stephan and Jan Sprenger. 2012. "Judgment aggregation and the problem of tracking the truth." *Synthese* 187:209-221.

Hauray, Boris and Philippe Urfalino. 2007. "Expertise scientifique et intérêts nationaux. L'évaluation européenne des médicaments 1965-2000". *Annales HSS* 2: 273-298.

List, Christian. 2012. "The theory of judgment aggregation: an introductory review." *Synthese* 187:179-207.

Lorenz, Jan. 2007. "Continuous Opinion Dynamics under Bounded Confidence: A Survey." *International Journal of Modern Physics C* 18, 1819.

Martini, Carlo and Jan Sprenger. Forthcoming. "Opinion aggregation and individual expertise." In *Scientific collaboration and collective knowledge*, ed. by T. Boyer-Kassem, C. Mayo-Wilson and M. Weisberg, Oxford University Press.

McConway, Kevin J. 1981. "Marginalization and Linear Opinion Pools." *Journal of the American Statistical Association* 76(374): 410-414.

Risse, Mathias. 2004. "Arguing for Majority Rule". *The Journal of Political Philosophy* 12(1): 41-64.

Stocker Thomas .F. et al. 2013. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Urfalino, Philippe and Pascaline Costa. 2015. "Secret-Public Voting in FDA Advisory Committees." In *Secrecy and Publicity in Votes and Debates*, ed. Jon Elster, 165-194. Cambridge University Press.

Wagner, Carl. 1982. "Allocation, Lehrer models, and the consensus of probabilities." *Theory and Decision* 14: 207-220.

Responsiveness and robustness in the David Lewis signalling game

Carl Brusse and Justin Bruner

October 28, 2016

Abstract

We consider modifications to the standard David Lewis signalling game and relax a number of unrealistic implicit assumptions that are often built into the framework. In particular, we explore realistic asymmetries that exist between the sender and receiver roles. We find that endowing receivers with a more realistic set of responses significantly decreases the likelihood of signalling, while allowing for unequal selection pressure often has the opposite effect. We argue that the results of this paper can also help make sense of a well-known evolutionary puzzle regarding the absence of an evolutionary arms race between sender and receiver in conflict of interest signalling games.

1 Signalling games and evolution

Common interest signalling games were introduced by David Lewis (Lewis, 1969) as part of a game theoretic framework which identified communicative conventions as the expected solutions to coordination problems. In recent years, this has informed a growing body of work on the evolution of communication, incorporating signalling games into an evolutionary game theoretic approach to modelling the evolution of communication and cooperation in humans (Skyrms, 2010; Skyrms, 1996).

As the basis for game theoretic modelling of such phenomena, David Lewis signalling games are attractive in their intuitive simplicity and clear outcomes. They are coordination games of common interest between world-observing senders and action-making receivers using costless signals; in contrast to games where interests may differ and where costly signals are typically invoked. In the standard two-player, two-state, two-option David Lewis signalling game (hereafter the '2x2x2 game'), the first agent (signaller) observes that the world is in one of two possible states (state1 or state2) and broadcasts one of two possible signals (signal1 or signal2) which are observed by the second agent (receiver) who performs one of two possible actions (act1 or act2). If the acts match the state of the world (i.e. act1 if state1 or act2 if state2) then the players receive a greater payoff than otherwise.

Most importantly, though, the game theoretic results are unequivocal. There exist two Nash equilibria that are, in Lewis's words, signalling systems where senders condition otherwise arbitrary signalling behaviour on the state of the world, and receivers act on those signals to secure the mutual payoff. The two

systems only differ on which signal gets to be associated with each state of the world[1]. Huttegger (2007) and Pawlowitsch (2008) have shown that under certain conditions a signalling systems is guaranteed to emerge under the replicator dynamics, a standard model of evolution to be discussed further in section 4.

Of course the degree to which Lewis' approach makes sense is the degree to which we have confidence in the interpretation and application of such a highly idealised model to the more complex target systems. The obvious worry is that by introducing more realistic features into the model one might break or significantly dilute previous findings on the evolution of signalling.

Not surprisingly, then, recent work on Lewis signalling games has investigated the many ways in which such de-idealizations could occur. Some deviations from the standard Lewis signalling game include: more and varied states of the world, the possibility of observational error or signal error, noisy signals, partial deviation in interest between senders and receivers, the reception of more than one signal, and so on. Many such concerns are dealt with favourably in Skyrms (2010), and in work by others. For example Bruner et al. (2014) generalizes beyond the 2x2x2 case and Godfrey-Smith and Martinez (2013) and Godfrey-Smith (2015) mix signalling games of common interest and conflict of interest. One complication of the Lewis signalling game (particularly important for our purposes) is that signalling systems are not guaranteed in the simple 2x2x2 case when the world is biased. In other words, when the probabilities of the world being in state1 or state2 are not equal, a pooling equilibrium in which no communication occurs between sender and receiver is evolutionarily possible.

2 Symmetry breaking

The focus here will be with the idealisation that sender and receiver are equally responsive in strategic settings. Senders and receivers (in the evolutionary treatment of such games) are two populations of highly abstract and constrained agency roles: all that signallers do on observing the state of the world is send a signal, and the receivers must act as though the world is in one or other of the sender-observable states. Of those two roles, it is the restriction on receivers which is the more problematic.

Imagine for example a forager sighting a prey animal at a location inaccessible to her, but close enough to be acquired by an allied conspecific (who cannot observe the animal). In this case, it is easy for the first forager to slip into the signalling role and execute it, whistling or gesturing to her counterpart. To play the receiver role, however, the second forager has to actually re-orient their attention (to some degree) and attempt to engage in appropriate behaviour for the world-state the first has observed (e.g. prey is to the east or to the west, etc.).

The Lewis signalling model by design is constrained such that the receiver's actions are limited to just those acts associated with the sender's observed world-states. It is of course sensible to begin inquiry with as simple of a model as possible and consider a limited range of responses to stimuli. However, our point is that it is more plausible to make these idealizations for signallers than

[1]The other two possible outcomes of the game are 'pooling equilibrium', where the receiver plays act1 or act2 unconditionally.

for receivers. Signals are (by stipulation) cheap and easy to send, yet the actions available to the receiver are less plausibly interpreted as intrinsically cheap and free of opportunity cost.

In addition, the informational states drawn on by sender and receiver are also likely to be very different. Any real-life sender's observation of a world state will likely inform their motivations ('we should catch that animal') to dictate a fairly clear course of action ('try to direct the other agent's behaviour'). But all the receiver gets is a whistle, gesture or other signal which (by stipulation) has no pre-established meaning. The experience of observing a strategically relevant state of the world will typically be richer and more detailed than that of observing a strategically relevant artificial signal. All this leads to two concerns. Firstly, asymmetries in the strategic situations are likely to exist between senders and receivers. Receivers are likely to have locally reasonable options available to them other than those relevant to signaller-observed states of the world, and their responsiveness to the strategic situation is therefore less satisfactorily modelled by the strictly symmetric payoff structures of standard signalling games. Call this the structural responsiveness concern.

Secondly, given the likely differences in informational states, goal-directness, workload and opportunity cost implications of sender and receiver roles, we can expect the mechanisms (cognitive and otherwise) which instantiate them to differ as well, quantitatively and qualitatively. This implies that we should not expect their update-responsiveness in any given game to be equal either. Yet the working evolutionary assumption is that senders and receivers update their strategies in an identical manner, modelled using either learning dynamics or replicator dynamics. Call this the evolutionary responsiveness concern.

3 Hedgehog strategies and update asymmetry

The first of these concerns might sound like an argument for abandoning coordination games and moving toward 'conflict of interest' or 'partial conflict of interest' models. However the issue is more specific than this.

The structural responsiveness concern provides parallel motivation to one of Kim Sterelny's (Sterelny, 2012) concerns about Skyrms (2010) use of the Lewis model. Sterelny asks whether the availability of 'third options' on the part of the receiver might undermine the evolution of signalling even when these third options are less valuable than the payoff for successful coordination. As part of a discussion of animal threat responses, he labels this a 'hedgehog' strategy – taking an action which pays off modestly, regardless of the state of the world. To make this concrete, hedgehogs often roll into a ball in response to predators. This is a stark contrast to the more sophisticated behaviour of vervets, who have specific responses to specific threats. Yet the optimal response a vervet takes to one threat – climb a tree when confronted by a leopard – may lead to total disaster when used in response to another threat, such as an eagle. Hedgehogs avoid such outcomes by 'hedging' unconditionally so as to secure a modest payoff. Translated to signalling games, such a gambit may, in many cases, be more attractive than attempting to respond optimally to a signal[2].

[2]It is worth noting here that the 'hedgehog' strategy in this Lewis signalling game is in many ways analogous to the risk dominant 'hare' response in stag hunt games. Playing hare instead of stag allows the agent to avoid disaster, but only guarantees the individual a

This compliments the structural responsiveness concern: receivers (especially) might have other options of value which will stand in competition to those assumed in the standard signalling game. Something like these hedgehog strategies are plausible departures from the idealisation and should be expected on the part of the receiver given a realistic demandingness of the role. The question is whether (as Sterelny suspects) including hedgehog strategies might undermine the robustness of evolution toward signalling systems.

Our second concern pertaining to evolutionary responsiveness parallels a well-known evolutionary hypothesis: the so-called Red Queen effect. In competitive relationships such as predator-prey or parasite-host, the Red Queen hypothesis states that species will be constantly adapting and evolving in response to one another just to "stay in the same place" (Van Valen, 1973). This should also be the case in competitive signalling situations – such as predator-prey signalling systems or courtship displays among conspecifics. Signallers and receivers come to not just update their strategies, but to do so at faster or slower rates depending on the nature of the strategic encounter they are entwined in[3].

It might seem that in David Lewis signalling games (as with games of common interest in general) the Red Queen effect should have no role to play. However any realistic interpretation of the Lewis signalling game makes it plausible to consider asymmetry in evolutionary responsiveness as likely, if not the norm. First, as argued, the precise cognitive mechanisms and procedures employed by senders and receivers are likely to be different. Different systems will admit to different degrees of plasticity and evolvability – and will have a different set of cross-cutting tasks and utilities that will place their own demands upon them. Quick and easy signalling responses will have different pathways of update and adaptation than the (typically) more complex set of systems which appropriate receiver responses require.

The consideration of multiple use or adaptive reuse also makes the Red Queen hypothesis salient: it is wildly implausible that entirely separate cognitive systems would evolve to deal with competitive signalling situations and coordination-style situations. Cognitive structures which underpin sender or receiver behaviour will likely be subject to evolutionary pressures from competitive as well as cooperative situations, and the responsive nimbleness of sender and receiver strategies is therefore not guaranteed to be the same. We should not assume that the evolution of sender and receiver strategies always proceeds at the same pace.

Finally, there is at least some evidence of a basic asymmetry between sender and receiver roles in the literature on great ape communication. For example, Hobaiter and Byrne (2014) stress the great sophistication and flexibility on the receiver side of Chimpanzee gestural communication, while Seyfarth and Cheney (2003) discuss about how greater inferential sophistication on the receiver side is a feature of many primate communication systems. While these findings do

mediocre payoff. Thus the issues and trade-offs associated with the hedgehog strategy are general concerns not confined to just the Lewis signalling games. Thanks to [name redacted for review] for helping us better see this connection.

[3]An example of two groups adapting and evolving at different rates can be found in Richard Dawkin's discussion of his famous Life-Dinner principle (Dawkins and Krebs, 1979). While we expect both predator and prey to adapt to each other, Dawkins claims the prey species will come to evolve at a faster rate than the predator species due to the different selection pressures exerted on both species. Failing to adapt quickly enough for the predator means going hungry for an extra day, while failing to adapt for the prey means death.

not directly support the structural and evolutionary responsiveness concerns, they show that real-life sender and receiver strategies (in our near biological cousins at least) exhibit important differences, suggesting cognitive asymmetries compatible with those concerns.

In summary then, there is reason to consider two structural modifications to the Lewis signalling game as especially salient to the issue of responsiveness: the addition of 'hedgehog' strategies for receivers, and differing rates of change in sender and receiver strategies.

4 The model

The evolutionary model we use as a basis for our analysis is the pure-strategy 2x2x2 David Lewis signalling game, with the two-population discrete-time replicator dynamics.

Exact components of the model include two states of the world (L and R), a world-observing signaller with two possible signals (V1 and V2), and a signal-observing receiver with two possible actions (AL and AR). If the receiver's action matches the state of the world, then both signaller and receiver get a fixed positive success payoff, otherwise their payoff is zero. Signallers and receivers both have four pure strategies available to them (see table 1).

$S1$	Signal V_1 if L and signal V_2 if R
$S2$	Signal V_2 if L and signal V_1 if R
$S3$	Signal V_1 always
$S4$	Signal V_2 always
$S5$	Act A_L if V_1 and act A_R if V_2
$S6$	Act A_R if V_1 and act A_L if V_2
$S7$	Act A_L always
$S8$	Act A_R always

Table 1: Signaller and receiver strategies in the standard 2x2x2 common interest signalling game.

For the evolutionary model, the proportions of the different strategies within sender and receiver populations are initially randomly generated. The fitness of each strategy at a time period t is determined by the composition of the opposing population and the payoff associated with each strategy pairing. The proportion of each strategy at play in the next time period $t+1$ is determined by the standard discrete-time replicator dynamics. For the sender population this is:

$$X_i(t+1) = X_i(t)\frac{F_i}{F_S}$$

where X_i is the ith sender strategy, F_i is the fitness of that strategy and F_S is the average sender strategy fitness. Likewise, for receivers:

$$Y_j(t+1) = Y_j(t)\frac{F_j}{F_R}$$

where Y_j is the jth sender strategy, F_j is the fitness of that strategy and F_R is the average receiver strategy fitness. This is repeated until the populations settle

into an evolutionarily stable arrangement. The update process is deterministic and no randomising or mutations are allowed.

5 Modifications and results

We introduce two novel modifications to this model. First, we add a 'hedgehog' action A_H for the receiver. Second, we allow the rate of generational change of senders and receivers to vary relative to one other. In addition, the bias of nature is also varied, and we investigate the effects these three departures from the Skyrms/Lewis idealisation have on the evolutionary stability of signalling equilibria.

Turning to our first modification, the receiver now has three possible actions upon observing the signal: A_L, A_R, and A_H. As before a success payoff of 1 is received by both players in the case that the receiver plays A_L while the world is in state L, or the receiver plays A_R while the world is in state R. A payoff of zero is received if A_L or A_R is played otherwise. A payoff of H is received unconditionally if the receiver plays A_H, where the value of H is between 0 and 1. The sender has four familiar pure strategies, whereas the receiver now has five (for simplicity we omit conditional strategies involving A_H).

To adapt the earlier forager story, we can imagine the sender and receiver as an egalitarian hunting party, and the game as a situation where the sender remotely observes the location of a valuable prey animal (left or right) and calls out to the receiver. The receiver is initially unable to observe the prey but can choose to go left or go right (catching the prey if they go in the matching direction), or alternatively to abandon the hunt in order to obtain a less valuable resource they do not need help from the sender to acquire (the hedgehog strategy). Varying the prior probability of the world is equivalent to it being in a situation where it is systematically more likely that the prey is to the left or the right.

In the simple unbiased 2x2x2 signalling game, one of the two signalling equilibria is guaranteed to be reached under the replicator dynamics. In our notation, these equilibria are S1-R1 and S2-R2. Increasing the bias of the world (i.e. making L more probable than R or vice versa) will undermine this, with an increasing proportion of populations instead collapsing to pooling equilibria. This will occur when there are initially few conditional signalling strategies in the sender population. In such situations, receivers do best to simply perform the act that is most appropriate for the more likely state of the world. The incentive for senders to adopt a signalling system then disappears and the community is locked into a pooling equilibrium.

Not surprisingly, we found a similar effect with the hedgehog strategy as values of H, the payoff for A_H, becomes significant. The hedgehog strategy R5 is an additional unilateral response, and is able to draw some initial populations away from the signalling equilibria when H is in excess of 0.5 (i.e., the average payoff for 'guessing'). This result, for an unbiased world, is illustrated in Figure 1[4].

[4]Note that the exact range of this effect, including the point at which the effect becomes significant and the y-intercept, are artefacts of the number of world-states and strategies in the model and therefore not general.

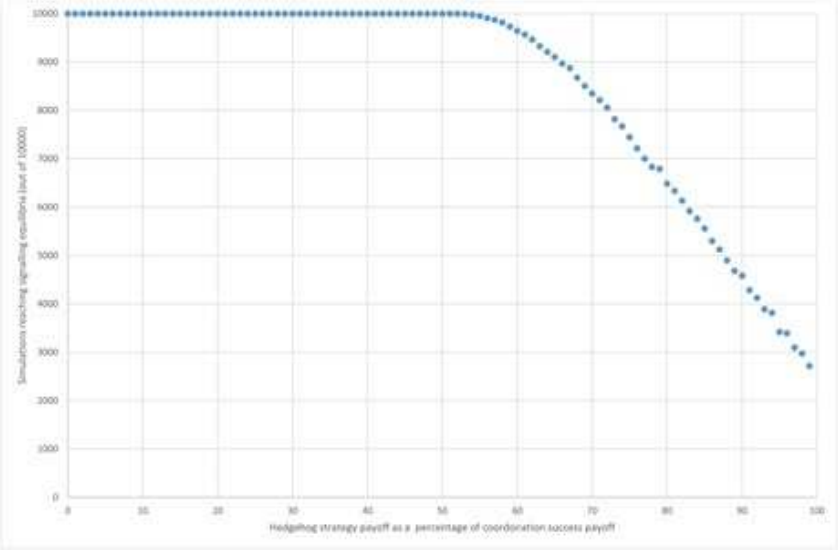


Figure 1: Effect of hedgehog payoff on proportion of signalling equilibria.

We observe a more surprising result when the bias and H are varied in combination. Figure 2 shows the results of varying bias for different values of H. The H = 0 curve has the expected n-shape, with perfect signalling being degraded as world-bias increases away from the mid-point of even bias between L and R. The inclusion of significant (i.e. H ¿ 0.5) hedgehog payoffs decreases signalling at even bias. As nature becomes increasingly biased, however, the proportion of simulations that head to a signalling system does not go down. In fact we observe a 'plateau' followed by a gradual *increase* in the proportion signalling as nature becomes increasingly biased. However, once the bias becomes too extreme, the traditional pooling equilibrium becomes increasingly likely as the payoff associated with simply performing the appropriate act for the more likely state of the world approaches 1. This results in a steep decline in the proportion of simulations that result in signalling systems.

6 Generational asymmetry

We now turn to our second modification of the David Lewis signalling framework in which we introduce a generational asymmetry. We introduced a 'slow-down factor' Z to the replicator dynamics in order control the rate at which sender and receiver populations change over time. Composition of the sender and receiver populations are now governed by the following equations:

$$X_i(t+1) = (1 - Z_S)X_i(t)\frac{F_i}{F_S} + X_i(t)Z_S$$

$$Y_j(t+1) = (1 - Z_R)Y_j(t)\frac{F_j}{F_R} + Y_j(t)Z_R$$

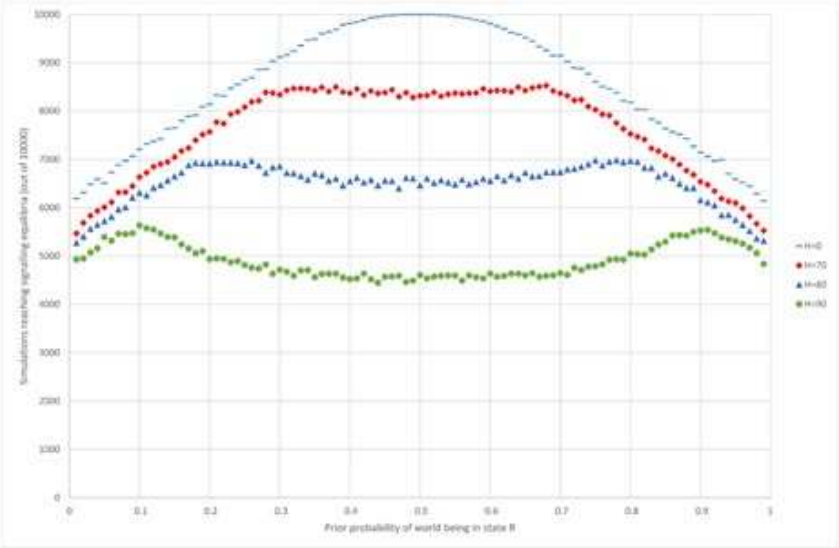


Figure 2: Effect of hedgehog strategy and bias of nature on proportion of signalling equilibria.

Note that when both Z_R and Z_S are zero there is no deviation from the standard replicator dynamics. Rates of changes are slowed as their values increase; for example setting $Z_S = .5$ halves the rate of change for sender strategies. Z_R (alone) being set to 1 means taht the composition of the receiver population would not change over time, and only the sender population would evolve.

The result of introducing this generational asymmetry between senders and receivers is that signalling is more likely when sender strategies evolve faster than receiver strategies. This is illustrated in figure 3, where senders (Z_S) and receivers (Z_R) are slowed down to half and one-tenth speeds (with the other population unaltered) as the bias of nature is varied.

Slowing the evolution of the sender population leads to more pooling because, as before, receivers facing a sender population whose conditional signalling is low will begin to gravitate to the act that matches the more likely state of the world (and the threshold for 'low' is higher at higher bias). This evolutionary trajectory only reverses if conditional signalling increases rapidly enough to tip the fitness balance toward its matching conditional response, before that response is overpowered. Thus signalling becomes quite a remote possibility when bias is high and senders are slow, occurring in less than 10% of simulations for some parameter values. Slowing the evolutionary responsiveness of the receiver population evolves has the opposite effect – as senders will have time to adopt the best separating strategy given the mix of receiver strategies, and the receiver population slowly adjusts and a robust signalling system establishes. By a similar logic, it is easy to see that a quickly evolving sender population also mitigates against the effect of hedgehog strategies.

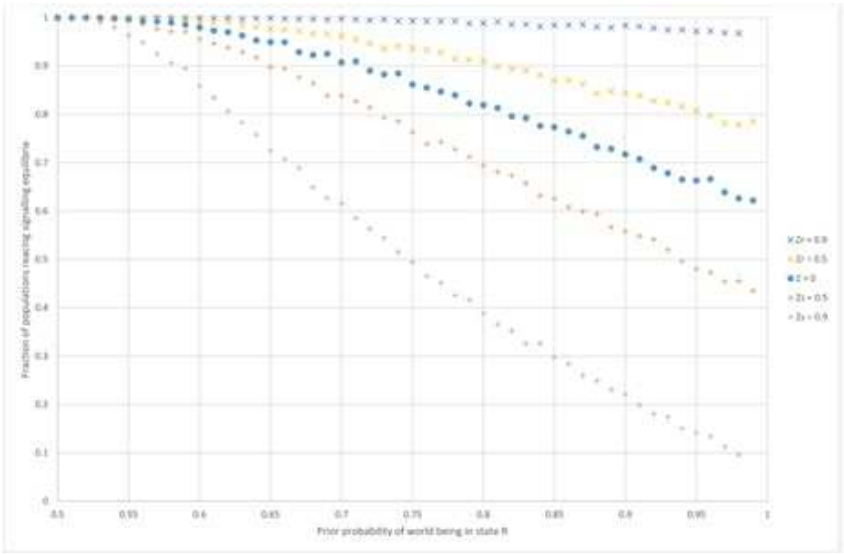


Figure 3: Effect of generational asymmetry and bias of nature on proportion of signalling equilibria.

7 Discussion

We have explored a few well-motivated departures from the highly idealized and simple Lewis signalling game typically considered in the literature. As shown in section 4, breaking the symmetry between senders and receivers often significantly reduces the likelihood that a separating equilibrium emerges. For one, providing receivers with a safe third option which allows them to secure a decent payoff regardless of the state of the world significantly reduces the size of the basin of attraction of the separating equilibrium. Likewise, separating is a remote possibility when receivers outpace senders in the race to adapt.

However the interaction between hedgehog payoffs and bias shows that signalling-undermining effects are not strictly additive. Likewise, the situation is much less bleak when senders evolve at a faster pace than receivers. Interestingly, many scholars in the animal communications literature have noted a similar response asymmetry between sender and receiver in conflict of interest and partial conflict of interest signalling games. For instance, Owren, Rendall, and Ryan (2010) note that senders can easily adapt their signalling behaviour while receivers for the most part have responses to the stimuli produced by senders that are more difficult to change. Thus some have taken to think of signalling as primarily involving the manipulation of receivers by senders.

But this leaves us with an evolutionary puzzle. If there is a conflict of interest between sender and receiver, then what prevents receivers from increasing the speed at which they adapt to the behaviour of the senders? In other words, what explains the absence of an evolutionary arms race between sender and receiver? These are the exact circumstances we would expect the red queen hypothesis to apply. We believe the results of this paper may form the basis of

a novel explanation for this puzzling phenomena. When the interests of sender and receiver are perfectly aligned it is actually in the interest of both parties for the sender population to 'take the lead' and evolve at the faster rate, as doing so ensures the community is more likely to hit upon a mutually beneficial signalling system. When the interests of sender and receiver significantly diverge, however, we would expect this not to be the case since both parties now have reason to adapt at a faster pace than the other.

Yet individuals who routinely interact rarely find themselves playing either common interest or conflict of interest signalling games exclusively. As is well known by any parent, not all signalling interactions between relatives are free of conflict. Likewise, agents whose interests are typically thought to be partially opposed, such as two potential mates, may frequently engage in common interest signalling games in contexts unrelated to mating. The point is that a variety of strategic scenarios can hold between sender and receiver, and there is no principled reason to think all interactions will involve perfect alignment or sizable conflict. If so, then a proportion of signalling interactions between sender and receiver may involve no conflict, a partial conflict, or a full conflict of interest. When the proportion of no or low conflict signalling games is significant, the generational asymmetry result from the previous section may hold to some degree. Both sender and receiver will then profit from the sender population evolving at a faster rate than the receiver population, and receivers do best to limit how responsive they are to senders so as to ensure the emergence of informative signalling systems when their interests do overlap. Thus, while it may appear puzzling as to why a receiver is not more responsive when her interests diverge from that of the sender, this confusion might be resolved when the interaction is put into context.

The robustness analysis considered in this paper has in some sense shown how fragile the evolution of signalling can be. Slightly altering the framework in a sensible fashion leads to significantly different results. While many variants of the baseline Lewis signalling game have been explored by philosophers in recent years, more work is required in order to better assess the prospect of signalling in realistic environments.

8 Acknoweldgements

We thank Kim Sterelny, Ron Planer and the audiences at the Sydney-ANU Philosophy of Biology Workshop and the 2016 Meeting of the Philosophy of Science Association.

9 Bibliography

Bruner, Justin, Cailin O'Connor, Hannah Rubin, and Simon M. Huttegger. 2014. "David Lewis in the Lab: Experimental Results on the Emergence of Meaning." Synthese, September, 1–19. doi:10.1007/s11229-014-0535-x.

Dawkins, R., and J. R. Krebs. 1979. "Arms Races between and within Species." Proceedings of the Royal Society of London B: Biological Sciences 205 (1161): 489–511. doi:10.1098/rspb.1979.0081.

Godfrey-Smith, Peter. 1996. Complexity and the Function of Mind in Nature. Cambridge; New York: Cambridge University Press.

Godfrey-Smith, Peter, and Manolo Martínez. 2013. "Communication and Common Interest." PLoS Comput Biol 9 (11): e1003282. doi:10.1371/journal.pcbi.1003282.

Hobaiter, Catherine, and Richard W. Byrne. 2014. "The Meanings of Chimpanzee Gestures." Current Biology 24 (14): 1596–1600. doi:10.1016/j.cub.2014.05.066.

Huttegger, Simon M. 2007. "Evolution and the Explanation of Meaning*." Philosophy of Science 74 (1): 1–27. Lewis, David K. 1969. Convention: A Philosophical Study,. Cambridge, MA: Harvard University Press.

Martınez, Manolo, and Peter Godfrey-Smith. 2015. "Common Interest and Signaling Games: A Dynamic Analysis." http://petergodfreysmith.com/wp-content/uploads/2013/06/Martinez-GS-paper2-Dynamic-Preprint.pdf.

Owren, Michael J., Drew Rendall, and Michael J. Ryan. 2010. "Redefining Animal Signaling: Influence versus Information in Communication." Biology and Philosophy 25 (5): 755–80. doi:10.1007/s10539-010-9224-4.

Pawlowitsch, Christina. 2008. "Why Evolution Does Not Always Lead to an Optimal Signaling System." Games and Economic Behavior 63 (1): 203–26. doi:10.1016/j.geb.2007.08.009.

Seyfarth, Robert M., and Dorothy L. Cheney. 2003. "Signalers and Receivers in Animal Communication." Annual Review of Psychology 54 (1): 145–73. doi:10.1146/annurev.psych.54.101601.145121.

Skyrms, Brian. 1996. Evolution of the Social Contract. Cambridge University Press. ———. 2010. Signals: Evolution, Learning, and Information. Oxford; New York: Oxford University Press.

Sterelny, Kim. 2012. "A Glass Half-Full: Brian Skyrms's Signals." Economics and Philosophy 28 (01): 73–86. doi:10.1017/S0266267112000120.

Van Valen, Leigh. 1973. "A New Evolutionary Law." Evolutionary Theory 1 (1-30). http://tmtfree.hd.free.fr/albums/files/TMTisFree/Documents/Biology/A

Experimental Individuation and Retail Arguments

Ruey-Lin Chen
Department of Philosophy, National Chung Cheng University, Taiwan

Jonathon Hricko
Education Center for Humanities and Social Sciences, National Yang-Ming University, Taiwan

Abstract: Magnus and Callender (2004) argue that we ought to focus on retail arguments, which are arguments regarding the existence of particular kinds of theoretical entities, as opposed to theoretical entities in general. However, *scientists* are the ones who put forward retail arguments, and it's unclear how *philosophers* can engage with such arguments. We argue that philosophers can engage with retail arguments by providing criteria that they must satisfy in order to demonstrate the existence of theoretical entities. We put forward experimental individuation as such a criterion—when scientists experimentally individuate an entity, a realist conclusion about that entity is warranted.

Word Count: 4983

1. Introduction

Magnus and Callender argue that we ought to abandon "wholesale arguments," which are "arguments about all or most of the entities posited in our best scientific theories" (2004, 321). Instead, we ought to embrace "retail arguments," which are "arguments about specific kinds of things such as neutrinos, for instance" (2004, 321). This shift in focus rules out standard scientific realism as well as various antirealist positions, and in Section 2, we'll argue that Magnus and Callender's position is preferable to these other positions.

However, we recognize that philosophers who choose to abandon wholesale arguments in favor of retail arguments face a potential problem. Dicken (2013) has argued that such philosophers will merely end up repeating the retail arguments that scientists offer. In that case, the turn to retail arguments may entail that no distinctively philosophical work remains to be done. In Section 3, we'll argue that this is not the case. Not all retail arguments successfully demonstrate the existence of theoretical entities, and it can take some philosophical work to distinguish the ones that do from the ones that don't.

In Section 4, we'll put forward a criterion for doing so, which we take from Chen's (2016) work on experimental individuation. Chen suggests that "[i]f a scientist can realize the individuality of an object in a particular experiment, then she has provided the strongest evidence ... to warrant the reality of the object" (2016, 365). We'll argue that retail arguments that demonstrate the experimental individuation of a theoretical entity succeed in showing that realism about that entity is warranted.

We'll draw on three examples throughout the paper: Lavoisier's oxygen theory of acidity, J. J. Thomson's work on cathode rays, and Davy's discovery of potassium. We'll conclude, in Section 5, by applying our criterion to these three cases, with the result that the upshot of a retail argument can be either realism, antirealism, or skepticism regarding the existence of a particular kind of theoretical entity.

2. The Turn to Retail Arguments

We'll now introduce Magnus and Callender's position in a bit more detail, and indicate why we take it to be preferable to standard scientific realism (SSR) and antirealism. SSR is a position regarding theories in general—the success of our best theories warrants the claim that they are at least approximately true, as well as the claim that the theoretical entities that they posit exist. Antirealist positions come in a number of different forms, but they all typically endorse claims about theories in general, and deny that success warrants the two claims endorsed by proponents of SSR.

According to Magnus and Callender, there is something that all of these positions have in common, namely, their proponents attempt to support these positions by engaging in wholesale arguments. They focus on two examples of such arguments. First of all, there is the no-miracles argument, according to which the success of our best theories would be a miracle if those theories weren't at least approximately true. Secondly, there is the pessimistic meta-induction, which uses past successful-but-false theories as an inductive basis for concluding that our current successful theories are false as well. The no-miracles argument is taken to support "[w]holesale realism," which "seeks to explain

the success of science in general"; and the pessimistic meta-induction is taken to support "wholesale anti-realism," which "seeks to explain the history of science in general" (2004, 321). However, Magnus and Callender argue that these arguments, and wholesale arguments in general, ought to be abandoned. This is because they embody the base rate fallacy, since they don't take into account the base rate probability of any successful theory being true or false. For this reason, they maintain that wholesale realism and wholesale antirealism ought to be abandoned as well.

Magnus and Callender propose that we ought to replace wholesale arguments with retail arguments. Unlike wholesale arguments, the scope of a retail argument is restricted to a particular theory and/or a particular kind of theoretical entity. By shifting the focus from theories in general to theories in particular, philosophers can *dissolve* the traditional realism debate, with the result that "realism and anti-realism are options to be exercised sometimes here and sometimes there" (2004, 337). This, in turn, opens up the possibility that "[t]here may be good reasons to be a realist about neutrinos, an anti-realist about top quarks, and so on" (2004, 333).

In order to show why this possibility represents an improvement over SSR and antirealism, we'll now consider a case from the history of chemistry. This case concerns the composition of hydrochloric acid. Scheele was the first to decompose this acid, which he called "acid of salt," and he identified its constituent substances as phlogiston and "dephlogisticated acid of salt" (1774/1931). However, it was a matter of some controversy whether he had succeeded in decomposing hydrochloric acid. According to Lavoisier's oxygen theory of acidity, all acids are composed of oxygen (the principle of acidity) and a radical, which can be either a simple substance or a compound (1789/1965,

65, 115). Neither Scheele nor any other chemist had been able to extract the oxygen from hydrochloric acid, which Lavoisier called "muriatic acid." And so Lavoisier held that it remained undecomposed, and, in accordance with his theory, he hypothesized that it must contain oxygen combined with what he called "the muriatic radical" (1789/1965, 71-72). As for Scheele's dephlogisticated acid of salt, Lavoisier held that it is a compound of muriatic acid and oxygen, which he called "oxygenated muriatic acid" (1789/1965, 73). Some years later, Davy argued that Scheele was correct, while Lavoisier was in error (1810, 236-37). On Davy's view, muriatic acid is composed of hydrogen and what he calls "oxymuriatic acid," which is what Lavoisier called "oxygenated muriatic acid," and what Scheele called "dephlogisticated acid of salt." Davy later went on to argue for the elementary nature of this latter substance, and proposed a new name for it: "Chlorine" (1811, 32). His approval of Scheele stems from the fact that Davy, like a number of latter-day phlogiston theorists, identified hydrogen with phlogiston.[1] And the claim that hydrochloric acid is made up of hydrogen and dephlogisticated acid of salt, even if terminologically problematic, is essentially correct. Lavoisier, however, was in error since this acid contains no oxygen, thus falsifying his oxygen theory of acidity.

Proponents of SSR, impressed by narratives of the Chemical Revolution according to which Lavoisier's oxygen theory defeated the phlogiston theory, are often explicit that their realism applies to the oxygen theory but not to the phlogiston theory.[2] But in that case, SSR entails the implausible conclusion that Lavoisier's muriatic radical exists, while Scheele's dephlogisticated acid of salt does not. It seems much better to

[1] See, e.g., Kirwan (1789, 4-5).

[2] See, e.g., Hardin and Rosenberg (1982, 610) and Psillos (1999, 291).

conclude that Lavoisier's muriatic radical doesn't exist, while Scheele's dephlogisticated acid of salt does.

Antirealism, at least of the Kuhnian variety, fares no better. Those influenced by Kuhn's (1962/1996) views regarding incommensurability would claim that theoretical entities conceptualized by rival theories should be treated as different entities. However, chemists working in the late eighteenth and early nineteenth centuries shared a set of operations for producing the substance that was variously known as dephlogisticated acid of salt, oxymuriatic acid, and chlorine. It's therefore implausible to maintain that, in light of the fact that these chemists held different theories, they were working with distinct theoretical entities. A trans-theoretical view of the substance that came to be known as chlorine is therefore preferable.

By abandoning wholesale arguments in favor of retail arguments, we can sidestep these difficulties, and simply adopt realism about chlorine (whatever it was called and however it was conceptualized) and antirealism about Lavoisier's muriatic radical. That said, by trading wholesale arguments for retail arguments, we face another difficulty, to which we'll now turn.

3. Can Philosophers Engage with Retail Arguments?

Dicken (2013) has objected that those who abandon wholesale arguments in favor of retail arguments face a serious difficulty. In short, once one does so, it's not clear that any "distinctively philosophical" issues remain to be addressed (2013, 564). Scientists are generally the ones who put forward retail arguments. And if the turn to retail arguments

amounts to merely repeating arguments scientists have offered first, then perhaps nothing distinctively philosophical remains to be done. Our goal in the remainder of the paper is to provide a way of engaging with retail arguments that is distinctively philosophical, and to thereby answer Dicken's objection.

We'll start by considering how scientists demonstrate the existence of theoretical entities, and so we'll now introduce another case from the history of science. This case concerns Thomson's work on cathode rays and his determination of the mass-to-charge ratio (m/e) of the electron. According to the official website of the Nobel Prize, it was because of this work that Thomson "received the Nobel Prize in 1906 for the discovery of the electron, the first elementary particle."[3] Thomson (1897, 1906/1967) hypothesized that cathode rays are currents of "carriers of negative electricity" or "corpuscles"—what we now know as electrons.[4] His hypothesis was not only about the nature of cathode rays, but also about the interaction among cathode rays and other theoretical entities such as electrostatic fields and electrons. In order to determine the mass-to-charge ratio, he measured the deflection of cathode rays passing through an electrostatic field, the strength of the electrostatic field, and other related magnitudes. He interpreted the value that he obtained for m/e in light of his hypothesis, and his experimental results confirmed that hypothesis.

---

[3] Retrieved January 27, 2016 from http://www.nobelprize.org/educational/physics/vacuum/experiment-1.html. See also Harré (2002) and Whittaker (1989).

[4] For the identification of Thomson's carriers with electrons, see the reprint of Thomson (1897) in Magie (1969), in which Magie makes the identification.

However, one might ask how it's possible to infer from Thomson's experimental confirmation of his hypothesis to the claim that he had thereby demonstrated the existence of the electron. Philosophers can engage with such a question. And regardless of the answers they provide, they must at least defend those answers by invoking some kind of criterion for concluding that the evidence that scientists have offered does or does not constitute a demonstration of the existence of a given entity. To take one example of such a criterion, Hacking (1983, 23) suggests manipulation: "if you can spray them then they are real." While Thomson manipulated cathode rays, he did not manipulate electrons, and so, according to Hacking's criterion, Thomson did not offer evidence strong enough to demonstrate the existence of electrons.

The important point, for our purposes, is that providing a criterion for granting the reality of a theoretical entity, and determining whether the evidence that scientists have offered satisfies that criterion, constitutes a way for philosophers to engage with retail arguments. Scientists may be the ones who initially put forward retail arguments. But it is a distinctively philosophical task to determine a criterion that can distinguish those retail arguments that demonstrate the existence of a theoretical entity from those that do not. We thus have a way of answering Dicken's objection, provided that, by invoking such a criterion, we are not thereby turning back to wholesale arguments. In the next section, we'll introduce our criterion and argue that applying it does not amount to a wholesale argument.

4. Ontological Commitment and Experimental Individuation

Our proposed criterion for granting the reality of theoretical entities is experimental individuation. A retail argument that demonstrates the experimental individuation of an entity is a good argument for realism about that entity.

Individuation and ontological commitment are connected. When scientists are ontologically committed to the theoretical entities that they posit, this commitment involves not just a belief that the entities exist, but also a responsibility to demonstrate their existence. Demonstrating the existence of a posited entity requires scientists to find an individual instance or sample of that entity, and if a scientist posits a theoretical entity without individuating it, then her ontological commitment is empty.

How do scientists individuate theoretical entities? Answering this question requires us to distinguish *theoretical individuation* from *experimental individuation*. Scientists theoretically individuate an entity if, in the course of theorizing, they describe a set of properties and behaviors of a posited entity by which they can identify it and distinguish it from other entities. However, these descriptions by which scientists theoretically individuate entities require evidence. Scientists can offer evidence for the existence of a theoretical entity if they produce an instance or sample of such an entity by performing an experiment. In doing so, they individuate an entity experimentally.[5]

The relationship between theoretical individuation and experimental individuation is much the same as the relationship between theory and experiment more generally.

---

[5] Scientists may also individuate an entity *observationally*, by observing an instance or sample of such an entity. Since observation is itself a complex issue, and since participants in the realism debate rarely question the existence of entities that scientists have observed, we will not discuss observational individuation here.

Various worries about the theory-ladenness of experimentation are relevant here. If a theoretical hypothesis yields a prediction regarding some experimental result, the result may be interpreted in light of the hypothesis. Moreover, since a theoretical hypothesis may involve two or more theoretical entities and their interactions, it can be difficult to show that an experiment produces an instance or sample of the target entity, i.e., that it experimentally individuates that entity. And it can be difficult to judge whether an experiment produces a real individual, as opposed to a mere phenomenon that results from experimental apparatuses and their interactions with experimented objects. For these reasons, a criterion of experimental individuation that is sufficiently independent of theoretical interpretation is needed.

Is there such a criterion for experimental individuation? One candidate is Hacking's manipulation criterion, which we mentioned in Section 3. However, since experimenters can manipulate not just real individuals, but also mere phenomena, manipulation cannot singly serve as the criterion of experimental individuation. Chen (2016) takes Hacking's criterion of manipulation, along with two other criteria, namely, separation and maintenance of structural unity, as jointly constituting a necessary and sufficient condition for the experimental individuation of a theoretical entity. In short, experiments that produce individuals are experiments that separate individuals from their surrounding environment, manipulate them, and maintain their structural unity throughout the process. Importantly, Chen's further conditions ensure that the manipulated object is a real individual as opposed to a mere phenomenon. We take Chen's criteria to offer a satisfactory account of experimental individuation. In Section 5, we'll illustrate his criteria in terms of three retail arguments from the history of science,

and thereby provide some support for our claim that his criteria are satisfactory.

For now, we wish to emphasize two points. First of all, experimental individuation is our proposed criterion for determining whether a retail argument successfully demonstrates the existence of some theoretical entity—it succeeds if it demonstrates the experimental individuation of that entity. Secondly, Chen's three criteria provide an adequate account of what experimental individuation requires.

Before moving on, we'll discuss two potential problems with this proposal. First of all, some theoretical entities, like the chemical substances named by mass terms like 'water,' 'phlogiston,' and 'oxygen,' are paradigm cases of non-individuals. It's therefore not immediately obvious how we can appeal to the notion of experimental individuation when it comes to such entities. We propose to do so by considering the experimental individuation of *samples* of such substances, as we'll illustrate in Section 5.1, in terms of Davy's discovery of potassium. Since samples count as individuals, our criterion is applicable to cases involving non-individuals like chemical substances.

Secondly, there's the issue as to whether the application of our criterion amounts to a kind of wholesale argument. Whether a given retail argument demonstrates the experimental individuation of some theoretical entity is a local matter, grounded in the details of that argument. In contrast, wholesale arguments are not grounded in such local matters. Instead, they rely on claims regarding populations of theories in general, and it is for this reason that they embody the base rate fallacy. We've consciously avoided reasoning that may lead to the base rate fallacy. For example, we haven't argued that the success of our best theories would be a miracle unless the entities they posit can be experimentally individuated. For these reasons, the application of our criterion to retail

arguments does not amount to a kind of wholesale argument. And in that case, we've provided a way of answering Dicken's objection, since our criterion provides a way for philosophers to engage with retail arguments.

5. Application of the Criterion to Three Retail Arguments

Our goal at this point is to show how one can use the criterion we've proposed in order to engage with retail arguments regarding the existence of particular kinds of theoretical entities. We'll discuss three cases: Davy's potassium, Lavoisier's muriatic radical, and Thomson's electron.

5.1 A Realist Conclusion Regarding Davy's Potassium

To begin with, we'll argue that Davy demonstrates the experimental individuation of potassium, and thereby provides us with a successful retail argument for realism about that substance.

Davy first isolated potassium by decomposing potash, which he did by means of electrolysis (1808, 4-5). He was the first to decompose potash, though for some time, chemists suspected it to be a compound.[6] Davy acted on a small piece of moistened potash with a Voltaic battery. As a result, at the negative surface of the battery Davy observed the appearance of "small globules having a high metallic lustre, and being precisely similar in visible characters to quicksilver" (1808, 5). In the lecture in which he

[6] See, e.g., Lavoisier (1965/1789, 156).

reports these results, Davy goes on to write: "These globules, numerous experiments soon shewed to be the substance I was in search of, and a peculiar inflammable principle the basis of potash" (1808, 5). And later in the lecture, he proposes the name "Potasium [sic]" for the basis of potash (1808, 32).

While this experiment, on its own, does not demonstrate the experimental individuation of a sample of potassium, subsequent experiments that Davy conducted do, and he shows that potassium satisfies all three of Chen's criteria. First of all, there is Chen's separation condition: scientists must separate the entities that they produce "from their environments" (2016, 348), and "from the experimental instruments that may have helped produce [them]" (2016, 365). In order to determine whether his results depended on the platinum instruments that he used, Davy performed a number of experiments using a variety of other materials, including copper, silver, and gold (1808, 5). And in order to determine whether his results depended on the fact that he conducted his experiments in the open atmosphere, he performed similar experiments in a vacuum (1808, 5). In all of these cases, he obtained the same results. These experiments collectively show that Davy had separated potassium from its surrounding environment (including the atmosphere and the other components of potash), and from the instruments that he used, thereby satisfying Chen's separation condition.

Secondly, there is Chen's condition regarding the maintenance of structural unity. Chen understands structural unity as the idea that "the components of an individual are structured into a whole in some specific manner" (2016, 358). Davy encountered a number of difficulties when it came to maintaining the structural unity of the globules of potassium that he had produced because "they acted more or less upon almost every body

to which they were exposed" (1808, 10). One of the first things Davy notes about the globules is that they did not last long—the ones that did not explode immediately after forming soon lost their metallic luster and became "covered by a white film" (1808, 5). Davy identifies this film as pure potash, and explains how it attracts moisture from the atmosphere, converting the globule into a saturated solution of potash (1808, 7). Eventually, Davy discovered one substance on which potassium did not have much of an effect, namely, recently distilled naphtha (1808, 10). He used that fluid to preserve globules of potassium, and he was able to examine the properties of potassium in the atmosphere by covering the globules with a thin film of naphtha. This method allowed Davy to maintain the structural unity of potassium, thus satisfying Chen's condition.

Thirdly, there is Chen's manipulation condition. Chen understands this condition in terms of the "instrumental use" of an object "to investigate other phenomena of nature" (2016, 358). Towards the end of the lecture in which he reports the electrolytic decomposition of potash, Davy conjectures that the globules of potassium he isolated "will undoubtedly prove powerful agents for analysis; and having an affinity for oxygene [sic] stronger than any other known substances, they may possibly supersede the application of electricity to some of the undecompounded bodies" (1808, 44). Making good on this conjecture would amount to showing that chemists can use potassium to decompose previously undecomposed substances, thereby satisfying Chen's manipulation condition. And in the following year, Davy made good on this conjecture by using potassium to extract the oxygen from a previously undecomposed substance, namely, boracic acid, thereby decomposing it (1809, 76-77).

In sum, Davy shows that samples of potassium satisfy all three of Chen's criteria.

And by demonstrating the experimental individuation of these samples, Davy presents us with a successful retail argument for realism about potassium.

5.2 An Antirealist Conclusion Regarding Lavoisier's Muriatic Radical

We'll now argue that Davy shows why the experimental individuation of Lavoisier's muriatic radical is not possible, and thereby provides us with a successful retail argument for antirealism about Lavoisier's radical.

As we discussed in Section 2, Lavoisier hypothesized that hydrochloric acid, which he called muriatic acid, is composed of oxygen and a hypothetical substance that he called the muriatic radical. He thereby theoretically individuated the muriatic radical as that substance which combines with oxygen to form muriatic acid, which, in turn, is converted into oxymuriatic acid (i.e., chlorine) by means of combining with even more oxygen. But as we emphasized in Section 4, theoretical individuation is a mere belief, and beliefs require evidence.

Davy (1810, 235-36) provides a retail argument that demonstrates that the experimental individuation of Lavoisier's radical is not possible. He emphasizes the results of various experiments that he and other chemists performed, which show that oxymuriatic acid combines with hydrogen to form muriatic acid. And he goes on to discuss those experiments that seem to show the decomposition of oxymuriatic acid into oxygen and muriatic acid. Davy observes that in these experiments, water is always present. And he concludes that the oxygen that such experiments produce results from the decomposition of the water, not from the decomposition of oxymuriatic acid, which has

not been demonstrated. If oxymuriatic acid doesn't contain oxygen, and muriatic acid contains oxymuriatic acid and hydrogen, then muriatic acid doesn't contain oxygen either. To adopt Davy's later terminology, the only components of muriatic acid are hydrogen and chlorine. Experimentally individuating the muriatic radical would involve separating it from the oxygen with which it combines to form muriatic acid and oxymuriatic acid. And since Davy showed that this is not possible, he gives us a successful retail argument for antirealism about Lavoisier's radical.

5.3 A Skeptical Conclusion Regarding Thomson's Electron

Finally, we'll argue that Thomson neither demonstrates the experimental individuation of the electron, nor shows that it is impossible. Hence, we have an example of an inconclusive retail argument. The proper response to such an argument is skepticism regarding the entity in question, at least until there is a conclusive retail argument regarding the existence of that entity.

Thomson (1897) designed a new type of cathode ray tube (figure 1) to perform a deflection experiment.

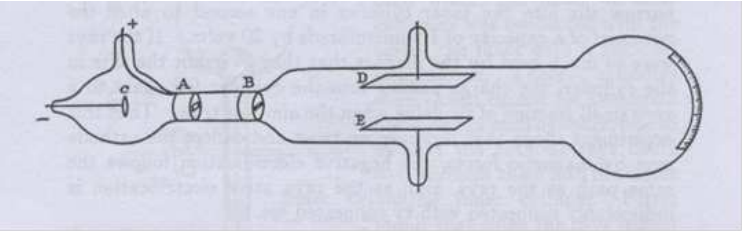


Figure 1. Thomson's cathode ray tube in 1897. Reproduced from Thomson 1969, 586.

This tube contains a cathode *C*, a cylindrical anode *A* with a slit, a cylindrical metal ring *B* with a slit, and a pair of plates *D* and *E* that produce an electrostatic field. A cathode ray is produced when the cathode discharges, and the ray passes through the slits in *A* and *B* before passing through the electrostatic field produced by *D* and *E*. Thomson's goal was to determine whether the ray would be deflected in the field, and to thereby determine the composition of cathode rays. The basic idea was that, if cathode rays were made of ethereal waves, the rays would not be deflected by an electrostatic field; if, however, the rays were made up of negatively electrified bodies, then the rays would be deflected by an electrostatic field.

Thomson's thought was that a cathode would produce both electric currents and cathode rays when discharging, and that, in order to determine the composition of cathode rays, it would be necessary to eliminate the electric currents and experiment with purified cathode rays. Purification is the function of the cylindrical metal ring *B*, which absorbs the electric currents leaked from *A* and thus ensures that the ray passing through *B* is pure. Thomson found that the purified cathode ray was deflected when it passed between the plates *D* and *E*, thus confirming that cathode rays are made up of negatively electrified bodies.

While Thomson satisfies Chen's criteria when it comes to cathode rays, he didn't thereby experimentally individuate the electrons that make them up. Thomson succeeded in *separating* cathode rays from currents; purifying them with the metal ring *B*, and thus *maintaining their structural unity*; and *manipulating* them by deflecting them with an electrostatic field. According to Chen's criteria, one can say that Thomson experimentally individuated cathode rays and demonstrated that they are currents of

negative electricity. But Thomson *presupposed* rather than demonstrated that the currents consist of electrons. He did not demonstrate the existence of electrons, because he did not experimentally individuate them. Hence, the proper response to the retail argument that Thomson gives us is neither realism nor antirealism, but rather skepticism regarding the existence of electrons, at least until there is a conclusive retail argument.

6. Conclusion

Our goal in this paper has been to provide a way for philosophers to engage with retail arguments, and thereby show that, even if we dissolve the traditional realism debate, there is still philosophical work to be done. We've put forward the criterion of experimental individuation in order to determine whether a given retail argument demonstrates the existence of a particular kind of theoretical entity. And we've applied that criterion to three cases, with the result that the upshot of a retail argument can be either realism, antirealism, or skepticism regarding the existence of a particular kind of theoretical entity.

References

Chen, Ruey-Lin (2016). "Experimental Realization of Individuality." In *Individuals Across the Sciences*, ed. Thomas Pradeu and Alexandre Guay, 348-70. New York: Oxford University Press.

Davy, Humphry (1808). "The Bakerian Lecture [for 1807], on Some New Phenomena of Chemical Changes Produced by Electricity, Particularly the Decomposition of the Fixed Alkalies, and the Exhibition of the New Substances Which Constitute Their Bases; And on the General Nature of Alkaline Bodies." *Philosophical Transactions of the Royal Society of London* 98: 1-44.

— (1809). "The Bakerian Lecture [for 1808]: An Account of Some New Analytical Researches on the Nature of Certain Bodies, Particularly the Alkalies, Phosphorus, Sulphur, Carbonaceous Matter, and the Acids Hitherto Undecompounded; With Some General Observations on Chemical Theory." *Philosophical Transactions of the Royal Society of London* 99: 39-104.

— (1810). Researches on the Oxymuriatic Acid, Its Nature and Combinations; And on the Elements of the Muriatic Acid. With Some Experiments on Sulphur and Phosphorus, Made in the Laboratory of the Royal Institution. *Philosophical Transactions of the Royal Society of London* 100: 231-57.

— (1811). The Bakerian Lecture [for 1810]: On Some of the Combinations of Oxymuriatic Gas and Oxygene, and on the Chemical Relations of These Principles, to Inflammable Bodies. *Philosophical Transactions of the Royal Society of London* 101: 1-35.

Dicken, Paul (2013). "Normativity, the Base-Rate Fallacy, and Some Problems for Retail Realism." *Studies In History and Philosophy of Science* 44(4): 563-70.

Hacking, Ian (1983). *Representing and Intervening*. Cambridge: Cambridge University Press.

Hardin, Clyde L. and Alexander Rosenberg (1982). In Defense of Convergent Realism. *Philosophy of Science* 49(4): 604-15.

Harré, Rom (2002). *Great Scientific Experiments: Twenty Experiments that Changed our View of the World*. New York: Dover.

Kirwan, Richard (1789). *An Essay on Phlogiston and the Constitution of Acids*. 2nd ed. London: J. Johnson.

Kuhn, Thomas S. (1962/1996). *The Structure of Scientific Revolutions*. 3rd ed. Chicago: University of Chicago Press.

Lavoisier, Antoine Laurent (1789/1965). *Elements of Chemistry*. New York: Dover.

Magie, William Francis, ed. (1969). *A Source Book in Physics*. Cambridge, Mass.: Harvard University Press.

Magnus, P. D. and Craig Callender (2004). "Realist Ennui and the Base Rate Fallacy." *Philosophy of Science* 71(3): 320-38.

Psillos, Stathis (1999). *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Scheele, Carl Wilhelm (1774/1931). On Manganese or Magnesia; and Its Properties. In *The Collected Papers of Charles Wilhelm Scheele, Translated from the Swedish and German Originals by Leonard Dobbin*, 17-49. London: G. Bell and Sons.

Thomson, Joseph John (1897). "Cathode Rays." *Philosophical Magazine, Fifth Series* 44: 293-316.

— (1906/1967). "Carriers of Negative Electricity. Nobel Lecture, December 11, 1906." In *Nobel Lectures: Physics, 1901-1921*, 145-53. Amsterdam: Elsevier Press.

— (1969). "The Electron." In Magie 1969, 583-97.

Whittaker, Edmund Taylor (1989). *A History of the Theories of Aether and Electricity*. New York: Dover.

Crash Testing an Engineering Framework in Neuroscience: Does the Idea of Robustness Break Down?[1]

ABSTRACT

In this paper I discuss the concept of *robustness* in neuroscience. Various mechanisms for making systems robust have been discussed across biology and neuroscience (e.g. redundancy and fail-safes). Many of these notions originate from engineering. I argue that concepts borrowed from engineering aid neuroscientists in (1) operationalizing robustness; (2) formulating hypotheses about mechanisms for robustness; and (3) quantifying robustness. Furthermore, I argue that the significant disanalogies between brains and engineered artefacts raise important questions about the applicability of the engineering framework. I argue that the use of such concepts should be understood as a kind of simplifying idealization.

"The brain is a physical device that performs specific functions; therefore, its design must obey general principles of engineering."

Sterling and Laughlin (2015:xv)

1. INTRODUCTION

In this paper I discuss a cluster of issues around the understanding of *robustness* in neuroscience. Systems biologist, Hiroaki Kitano defines

[1] M. Chirimuuta. History & Philosophy of Science, University of Pittsburgh. mac289@pitt.edu. Accepted for presentation at the 2016 Philosophy of Science Association meeting and publication of the proceedings in *Philosophy of Science*.

robustness as, "a property that allows a system to maintain its functions against internal and external perturbations" (Kitano 2004, p.826). According to this definition, in order to determine whether or not a system is robust, one must specify its function, and also specify the kinds of perturbation it faces. Empirically determinable questions then follow about how exactly the system achieves its robustness. Various means for making systems robust have been discussed across biology and neuroscience: copy redundancy, fail-safes, degeneracy, modularity, passive reserve, active compensation, plasticity, decoupling, and feedback (see Figure 1). It is obvious, but still worth emphasising, that most of these notions originate from engineering.

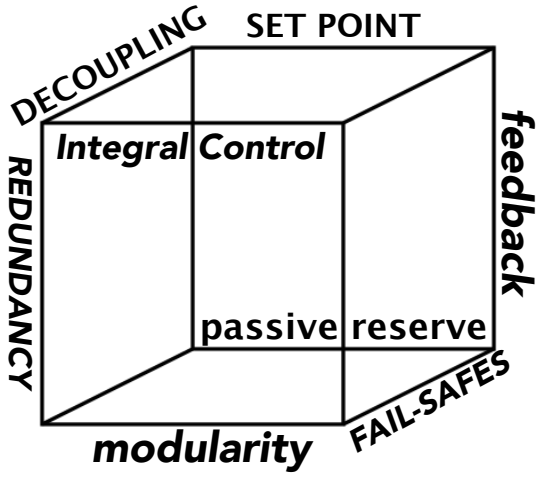


FIGURE 1. The Engineering Framework for Robustness. A set of terms originating from engineering and control theory, which are applied to biological systems to explain how they achieve robust performance.

In Section 2 of this paper I argue that the framework of concepts borrowed from engineering aids neuroscientists in (1) operationalizing robustness by specifying functions of the system and determining possible sources of

perturbation; (2) formulating hypotheses about means for the system to achieve robustness; and (3) showing how robustness may be precisely quantified. This will be shown with examples of neuroscientific research which aims to measure robustness in a retinal circuit (Sterling and Freed 2007), in the motor cortex (Svoboda 2015), and to develop models of homeostatic control (Davis 2006, O'Leary 2014).

In Section 3 I argue that the use of the engineering framework in neuroscience gets stretched, perhaps to breaking point, when applied to systems where (1) there is no principled distinction between processes for robustness and processes which continually maintain the life of the cell; (2) where perturbations are a regular occurrence rather than anomalous events; and (3) where one should not conceive of the system as seeking to maintain a steady state. This point will be illustrated through examination of some recent work from Eve Marder's laboratory, one of the key centres for research on robustness in neuroscience.

I will argue that the limitations of the engineering notions are put into stark relief when one examines neural systems through the lens of the process approach to biology (Dupré 2012). The engineering perspective, to the extent that it treats biological systems as pre-specified objects with fixed functions, misses many of the features that make robust biological systems fascinating and which are highlighted by the process view.

In Section 4 I will consider if it is necessary to re-engineer the concepts of robustness to be more in line with the dynamicism of biological systems; or alternatively, if we should accept the engineering perspective as it is, as one amongst many idealizing and simplifying heuristics for understanding complex systems like the brain.

2. PUTTING THE ENGINEERING FRAMEWORK TO USE

The robustness of the brain is one of its many extraordinary attributes. By this I mean the fact that brains can undergo moderately severe external perturbations while still maintaining approximately normal function. Obviously, robustness has its limits and the brain's characteristic patterns of resilience and fragility are an important target of research (Sporns 2010, chap. 10). In order to investigate robustness it is necessary first to specify what sorts of perturbations the system is robust to, and then to quantify how robust it actually is. Explanations of robustness can be developed by testing hypotheses concerning the exact mechanisms by which robust performance is achieved. The engineering framework can be put to effective use in each of these processes.

For example, Sterling and Freed (2007) pose the question of how robust the retinal circuit is. They define robustness as the factor by which intrinsic capacity exceeds normal demand, which is the engineer's notion of margin of safety (p.563). The idea can be illustrated through their comparison with bridge design. An engineer designing a road bridge will consider both the anticipated normal demand (e.g. commuter traffic) as well as the unusual demands that might occasionally be placed on the bridge (e.g. the passage of a 30 ton military vehicle). The unusual demand can be thought of as a "perturbation" in Kitano's terms. A robust design will ensure that the system does not break when pushed beyond normal conditions. For a bridge this can be achieved with passive reserve (using thicker steel than is needed under normal conditions) and redundancy (including additional beams so that there are back-up structures if any parts are compromised).

Sterling and Freed take the bridge case to be analogous to the retinal circuit. Normal demand, for the retina, is the intensity of illumination that the eye will encounter under naturalistic stimulation conditions. The safety factor is calculated by experimental determination of the maximum illumination level under which neurons in the retina can maintain their ability to signal to downstream neurons. Sterling and Freed (2007, p.570) report that,

> "across successive stages in this neural circuit, safety factors are on the order of 2–10. Thus, they resemble those in other tissues and systems. Their similarity across stages also accords with the principle of symmorphosis—that efficient design matches capacities across stages that are functionally coupled…."

Sterling and Freed's explanation of robustness depends on the notion of passive reserve. For photoreceptor neurons, this is calculated as the number of vesicles of neurotransmitter available in their synapse for continuous signalling at high-rates without restocking of the vesicles (p.565-6). In arriving at their conclusion about retinal safety margins, they argue that there are at least twice as many vesicles as needed under normal stimulation conditions. In this case we have seen that a design approach borrowed from civil engineering plays a clear and striking role in these neuroscientist's definition, operationalization and explanation of robustness in the retina.

Another example comes from Davis's (2006) review of work on homeostatic regulation[2] in the nervous system. As he writes:

[2] Note that Davis makes a conceptual distinction between robust properties and properties under homeostatic control: "In general, robustness describes a system with a reproducible output, whereas homeostasis refers to a system with a constant output" (2006, p.308). I will ignore this difference for the purposes of the paper since homeostatic systems conform to Kitano's general definition robust systems.

> "Homeostatic control systems are best understood in engineering theory, where they are routinely implemented in systems such as aircraft flight control. Recently, biological signaling systems have been analyzed with the tools of engineering theory…." (p.314)

Accordingly, homeostatic control systems have a number of "required features": 1) a set point which defines the target output of the system; 2) feedback; 3) precision in resetting the output back to the set point, following a perturbation; and (normally) 4) sensors which measure the difference between the actual output and the set point (p.309).

Thus control theory offers neuroscientists clear and experimentally testable criteria for determining whether a system undergoes homeostatic regulation, by looking for these required features (e.g. the existence of a set point) in a system. The operating conditions of homeostatic regulation, and the biophysical mechanisms of feedback, sensors, etc., are also open to experimental investigation. Reported examples of properties under homeostatic control are muscle excitation at the neuromuscular junction (p.309) and bursting properties of invertebrate neurons (p.311). More recently, O'Leary et al. (2014, p.818) argue that ion channel expression in their simplified model of invertebrate neurons can be understood as an implementation of *integral control*, a standard control-theoretic architecture.

Figure 2 (if space) schematic for integral control

3. CRASH TESTING THE FRAMEWORK

Before considering the question of whether the engineering framework becomes structurally unsound when applied to some kinds of neural systems, I would like to draw our attention to some of its features. The basic ideas are clearly illustrated in Sterling and Freed's (2007) example of the bridge. When one considers the robustness of an engineered artefact like the bridge, it is presupposed that the system is built up from component parts in such a way as to achieve a specific function. The robustness of the bridge is conceptually distinct from its other designed features or functions, and it can trade off against some of them. For example, the more robust the bridge is to the passage of the occasional heavy vehicle, the more expensive it will be to build (because requiring more steel) (p.563). Moreover, the perturbations against which the system is robust are thought of as atypical events, also conceptually distinct from the normal operations of the system.

There is also the tendency to think of robustness as allowing the system, following a perturbation, to return to its initial stable state. Some experiments specifically involve the operationalization of the robustness of a system as the reversion to a prior state. For example, reporting on an experiment in which mouse premotor cortex in one hemisphere was inhibited using optogenetics during the preparation period for the animal's movement, Svoboda (2015)[3] writes, that "[t]his preparatory activity is remarkably robust to large-scale unilateral optogenetic perturbations: detailed dynamics that drive specific future movements are quickly and selectively restored by the network." This notion of robustness as the ability of the system to revert to a

[3] To my knowledge, these results have not yet been published in a journal. I have contacted the author to find out if the study is under review or in press.

prior functional states is similar to the idea of *homeostasis* as the ability of a system to stabilize some quantity in spite of external changes.

Figure 3 (if space) After Kitano (2004, Figure 1)

Eve Marder's laboratory has carried out a long term investigation into the ability of neurons to maintain stable electrophysiological properties despite continual turnover of the ion channels embedded in the cell membrane which are responsible for its electrical excitability. This research project is one of the central examples of the study of robustness in neural systems. Marder and her collaborators make ample use of the engineering framework when reviewing other results and reporting their findings. For example, O'Leary et al. (2013, p.E2645) write:

> "Both theoretical and experimental studies suggest that maintaining stable intrinsic excitability is accomplished via homeostatic, negative feedback processes that use intracellular Ca^{2+} concentrations as a sensor of activity and then alter[s] the synthesis, insertion, and degradation of membrane conductances to achieve a target activity level."

What is striking about the characterization of electrophysiological stability in the face of ion channel turnover as a kind of robustness in the face of a perturbation (e.g. p.E2651), is the fact that the turnover is just part of the normal physiology of the cell. There is no functional and stable state of the cell in which this turnover does not occur—a fact which these authors also highlight.[4] This brings our attention to some strains in the application of the engineering framework to this biological system.

[4] "neurons in the brains of long-lived animals must maintain reliable function over the animal's lifetime while all of their ion channels and receptors are replaced in the membrane over hours, days, or weeks. Consequently, ongoing turnover of ion channels of various types must occur without compromising

In the basic engineering characterisation of robustness, sketched above, perturbations are different from the normal circumstances in which the system is expected to operate. "Perturbation" carries the everyday connotation of an event which throws the system off balance and is deleterious to its normal functioning. We cannot think of the events of ion channel turnover as perturbations in this sense; they are business as usual for the cell.

Furthermore, it is not in the nature of the system to seek to return to a prior, stable arrangement of its parts. A crucial property of the nervous system is its plasticity: the tendency for its component parts and the connections linking them to be continually sculpted by experience. The homeostatic mechanisms which Marder and colleagues investigate need to be understood as maintaining specific properties (such as a cell's Ca^{2+} concentration) at a certain point, but not (nor do these researchers claim it) some generalised operation for achieving system-wide internal stability (see §4.4).

In the basic engineering conception of robustness, there is a clear conceptual distinction between the features of a system which allow it to carry out its intended function, and those which make the system robust (even if in reality one individual feature can serve both purposes). In the case of the neuron which has continual ion channel turnover and no definite stable state to return to following these "perturbations", it is not clear that we can we make this distinction. A more natural way to think about this and other biological systems is as ones, unlike engineered artefacts, "designed" to keep changing

---

the essential excitability properties of the neuron" (O'Leary et al. 2013, p.E2645).

and "designed" to maintain functional stability in the midst of this constant change.[5]

The tensions and strains associated with the application of the basic engineering framework to biological systems can be felt more sharply if we appeal to a process metaphysics of biological "things" (Dupré 2012). According to this view, organisms are not substances but *processes*—items whose existence depends on the taking place of certain changes. This highlights the fact that the life of organisms depends on a continual turnover of its component parts, and that the system as a whole, while living, persists longer than its parts. Yet features and functions of the organism remain relatively stable. For example, memories can endure for decades even though the neurons that form them have undergone material change. This stability must be achieved—somehow. And so processes for robustness are not cleanly distinct from the general maintenance processes which keep the organism alive.

The processual nature of neurons is nicely described by Marder and Goaillard (2012, p.563):

> "each neuron is constantly rebuilding itself from its constituent proteins, using all of the molecular and biochemical machinery of the cell."
> (and see F n 4)

[5] This blurring of the lines between mechanisms for robustness and mechanisms for life is highlighted by Edelman & Gally (2001: 13763) in their discussion of the difference between redundancy and degeneracy in biological systems: "the term redundancy somewhat misleadingly suggests a property selected exclusively during evolution, either for excess capacity or for fail-safe security. We take the contrary position that degeneracy is not a property simply selected by evolution, but rather is a prerequisite for and an inescapable product of the process of natural selection itself." They also discuss another disanalogy between engineered and biological systems—the applicability of "design" talk.

We can contrast this with the substance metaphysics that we usually assume when thinking about engineered artefacts. A bridge or an aeroplane is what it is because of the parts which comprise it. Its existence does not depend on the occurrence of any process. This is not to deny that an expert in the theory of matter might well argue that the steel of the bridge maintains its integrity because of some fundamental processes. The point is that when characterising the robustness of the bridge or the aeroplane we would not resort to such sophistication. Rather, we think of the bridge as a substance and not a process—a steel structure which, in order to maintain its function in the face of perturbation, must resist rather than effect the swapping around of its component parts.

4. EXAMINING REASONS TO RE-ENGINEER

Now that we have noted these disanalogies between biological organisms and engineered things, we ought to worry that the framework borrowed from engineering is misleading when thinking about robustness in the brain and other biological systems. *Is it time to re-engineer our conceptual tools for thinking about robustness to make them more suitable for characterising living things?* In this section I consider four possible answers to this question.

4.1 *No. The terms in the engineering framework are just words that are used to facilitate communication of the neuroscientific results.*[6]

One potential response to the concerns raised in the previous section is that they stem from a superficial fixation on the vocabulary neuroscientists use when writing about their research. Just because the authors discussed above

[6] A response along these lines was suggested to me by Timothy O'Leary, in conversation.

have employed certain words first introduced by engineers, it does not follow that their understanding of neurophysiology is distorted by comparisons with engineering. For example, I mentioned that the word "perturbation" has a negative connotation which makes it seem inappropriate when describing non-pathological and frequent events like ion channel turnover. It could well be that in the context of this research the term takes on a different meaning—for example, as any event that the system cannot directly control,[7] such as changes in protein configuration due to thermal noise.

I believe that this response is warranted by what we know of the methodology of some of the investigations discussed above, but not all of them. In the case of Sterling and Freed (2007) I was careful to show that the engineering conceptions directly shaped how these neuroscientists operationalized and quantified robustness, and how they identified mechanisms by which robustness is achieved. There is no indication that they used terms such as "safety factor" to mean something radically different in the context of neuroscience.

A very explicit statement of the aim to apply engineering principles directly to the understanding of the premotor cortex comes from Svoboda (2015):

> "preparatory activity is distributed in a redundant manner across weakly coupled modules. These are the same principles used to build robustness into engineered control systems. Our studies therefore provide an example of consilience between neuroscience and engineering."

Thus the convergence between a neurophysiological and the engineering perspective on the mouse motor planning system is taken to be an important result of this study. This echoes Sterling and Laughlin's (2015, pp. xiii-xv) proposal that enquiring to see how engineering principles are implemented in

[7] I thank Timothy O'Leary for this suggestion.

neural systems, and the attempt thereby to reverse-engineer the brain, leads to insights not otherwise available through routine data collection.

4.2 *No. The inadequacies you point out with the engineering framework are based on a caricature of mechanical engineering, not the actual complex discipline.*[8]

My characterisation of the engineering framework assumes that mechanical engineering (the design of bridges, aeroplanes and such like) is paradigmatic of the engineering approach in general. But of course there are many different kinds of engineering, from mechanical to electronic to communications and chemical. It could well be that the mismatch between understanding the robustness of a highly dynamic entity like the brain, and the rather static conception of robust objects that falls out of the basic engineering framework is just an artefact of only focussing narrowly on the kind of engineering that is actually furthest away from neuroscience.

It would take me beyond the scope of this short article (and well beyond my own knowledge of the subject) to sketch out the various possible frameworks associated with each field of engineering specifically, and to see which conception of robustness is most suitable for biology. However, what I will say is that there is evidence in the studies discussed above that neuroscientists themselves do sometimes draw from the mechanically based caricature. This is particularly true of Sterling and Freed (2007). In contrast, when Davis (2006) and O'Leary (2014) make direct appeal to engineering they refer specifically to models in control theory.[9] This invites questions, still, about whether the paradigm examples of controlled systems (e.g. a car driven on

[8] This concern was raised by Arnon Levy and Timothy O'Leary.
[9] See also Zhang and Chase (2015) on the physical control system perspective on brain computer interfaces for motor rehabilitation.

cruise control, a Watt governor, or an aeroplane flown on autopilot) are dynamical enough capture the processual nature of the nervous system.

4.3 *Yes. The brain is so different from an engineered artefact that the framework is misleading and inappropriate.*

In Sections 4.1 and 4.2 I discussed two reasons for thinking that we should not be concerned about any radical disanalogy between robustness in biological and engineered systems. While I agree that these are important points to keep in mind, I do not think that they diffuse the fundamental concern that when neuroscientists borrow engineers' terms in order to study robustness, they risk mischaracterising the brain as more like an engineered artefact than it actually is. Is the appropriate conclusion, then, that a neural circuit is so different from a bridge or an aeroplane that the engineering framework is simply misleading and should be discarded?

The best way to make this strong negative case is to consider some historical examples in which reasoning by analogy with engineered systems seems to have lead neuroscientists and theorists astray. One example comes from von Békésy, a physicist and communications engineer who turned his attention to inhibition in the nervous system. In his book *Sensory Inhibition* he notes that there are feedback loops everywhere in nervous system and he asks how it is that system manages to avoid ending up in a dysfunctional oscillatory state (1967, p.25). It seems that von Békésy is importing his understanding of systems containing feedback from engineering, and in that context oscillations are normally problematic and efforts must be made to dampen them. These days neuroscientists seek to understand how oscillations in the healthy brain (i.e. its characteristic patterns of endogenous activity) are actually responsible for cognitive functions, and how these oscillations differ

from the ones associated with pathologies such as epilepsy and Parkinson's disease.[10]

Another example is the comparison of the effects of "noise" in brains and artificial signalling systems…… GET EXAMPLE

This is very different from how neuroscientists understand noise today, which begins with the idea that brains evolved under constraints imposed by noisy "components", which has therefore shaped all aspects of neural computation (Faisal et al. 2008). It would be a mistake to think of the brain processing information in the same way as an electronic computer, but with added redundancy to offset the noisiness of individual processing streams.

The cautionary tales just told give some concrete indications of how imposition of the engineering framework on to neural systems can lead to conclusions which in retrospect appear false and misguided. But it would be too hasty infer from these two examples that current work on robustness in neuroscience is of dubious standing whenever it appeals to the concepts of engineering. A more general argument is the following: *the brain is not like a bridge (or a computer, or an aeroplane on autopilot….); therefore whenever neuroscientists appeal to terms borrowed from the analysis of such systems, they risk saying things that are simply false because they fail to notice relevant disanalogies.* This lays all the sceptical cards on the table. In the last part of the paper I attempt to mitigate these worries.

[10] For a scientific overview see Buzsáki (2006). For discussion of philosophical implications, see Bechtel and Abrahamsen (2013). See also Knuuttila and Loettgers (2013, p.160) on a parallel difference across engineering and cell biology, where oscillations are found to have a functional role.

4.4 *No. Use of the engineering framework should be thought of as a simplifying strategy.*

Neuroscientist Steven Rose (2012:61) writes that:

> "one of the most common but misleading terms in the biology student's lexicon is homeostasis….[the] concept of the stability of the body's internal environment. But such stability is achieved by dynamic responses; stasis is death, and homeodynamics needs to replace homeostasis as the relevant concept"[11]

This seems to capture the problem that was first noted in Section 3, that we should not be mislead by the engineering framework into thinking of neural systems as seeking to maintain an initial stable state. But we also noted that the neuroscientists employing control-theoretic models of homeostatic mechanisms are not thinking of their systems as seeking stability in this very general way. Instead, they are modelling the stability of a specific variable—in the case of O'Leary et al. (2014), the concentration of Ca^{2+}—and investigating the mechanisms by which it is controlled. To this end, it is reasonable to interpret the system as an integral controller (p.818).[12] Thus it is still useful to talk about homeostasis with respect to Ca^{2+} concentration, even while thinking of the system as a whole, and in reality, as a "homeodynamic" one.

---

[11] Compare Sterling (2012) on the concept of *allostasis* – stability through change with an emphasis on predictive regulation. Day (2005) and O'Leary and Wyllie (2011), in contrast, argue that the concept of homeostasis easily accommodates these dynamic and predictive aspects, and that the term *allostasis* is therefore superfluous. It is an interesting question (but beyond the scope of this paper) whether the narrow or wide definition of *homeostasis* is currently more prevalent amongst biologists and neuroscientists.

[12] Note that O'Leary et al. (2014) study of homeostasis is via a *model* of a neuron. But the model is realistic enough that it is expected to shed light on actual biophysical mechanisms.

I think of neuroscientists whose investigation of robustness in the brain is scaffolded by the engineering framework as providing *idealized mechanistic explanations.* Their explanatory target is, for example, the process by which overall neuronal activity level is controlled via regulation of ion channel gene transcription through a Ca^{2+} sensitive feedback loop. This is standard fodder for mechanistic explanation. At the same time, the framework of engineering—in this case the schematic of the integral controller—serves to direct attention to specific parts and processes in the extremely complex cellular machinery and to interpret them in control theoretic terms (sensors, feedback loops, etc.), while bracketing other aspects not immediately relevant to the explanation of robustness.

Bechtel (2015, p.92) has presented the case that:

> "mechanisms are [to be] viewed not as entities in the world, but as posits in mechanistic explanations that provide idealized accounts of what is in the world."

His example is the idealization (understood as "falsehood") that scientists introduce by putting boundaries around putative mechanisms which in nature do not exist. In the cases explored in this paper, the idealization comes in through the analogical reasoning of treating a neuronal system *as if* it is an engineered artefact. This, like the positing of boundaries, is a useful way to simplify the explanandum. It enables neuroscientists to bracket some of the known facts about the brain's messy, Heraclitean nature. But it means, perhaps, that there is a stark difference between the brain viewed *sub specie aeternatis* (what some neuroscientists call the "ground truth" of the brain) and viewed *sub specie mechinae* (in the guise of a machine).

ACKNOWLEDGEMENTS

I am greatly indebted to Timothy O'Leary, Nancy Nersessian and Peter Sterling for their feedback on this work. I would also like to thank the

participants of the Fall 2015 workshop on Robustness in Neuroscience for discussion of the ideas behind this paper, and the audience at the Spring 2016 Re-Engineering Biology conference for their questions and comments on it. Both of these events were hosted by the Philosophy of Science Center at the University of Pittsburgh.

REFERENCES

Bechtel, W. and Abrahamsen, A. (2013). Thinking dynamically about biological mechanisms: Networks of coupled oscillators. *Foundations of Science*, 18:707–723

Bechtel, W. (2015). Can mechanistic explanation be reconciled with scale-free constitution and dynamics? *Studies in History and Philosophy of Science Part C,* 53: 84–93.

von Békésy, G. (1967). *Sensory Inhibition*. Princeton, NJ: Princeton University Press.

Buzsáki, G. (2006). *Rhythms of the Brain.* Oxford: Oxford University Press

Davis, G.W. (2006). Homeostatic control of neural activity: from phenomenology to molecular design. Annu. Rev. Neurosci. *29*, 307–323

Day TA (2005). Defining stress as a prelude to mapping its neurocircuitry: no help from allostasis. *Prog Neuropsychopharmacol Biol Psychiatry* **29**, 1195–1200

Dupré, J. (2012) *Processes of Life.* Oxford: Oxford University Press

Edelman GM, Gally JA (2001) Degeneracy and complexity in biological systems. Proc Natl Acad Sci USA 98(24):13763–13768.

Faisal, A., L. P. J. Selen and D. M. Wolpert (2008) Noise in the Nervous System. *Nature Reviews Neuroscience.* 9:292-303.

Kitano, H. (2004) Biological robustness. *Nature Reviews Genetics.* 5: 826-837.

Knuuttila, T. and A. Loettgers (2013). Basic science through engineering? Synthetic modeling and the idea of biology-inspired engineering. *Studies in History and Philosophy of Science, Part C 48*, 158–169.

Marder, E. and Goaillard, J.-M. (2012) Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*. 7:563-574

von Neumann, J. (2000). *The Computer and the Brain*. New Haven: Yale University Press.

O'Leary T, Williams AH, Caplan JC, Marder E (2013) Correlations in ion channel expression emerge from homeostatic tuning rules. *PNAS.* 110(28): 809–821

O'Leary T, Williams AH, Franci A, Marder E (2014) Cell types, network homeostasis and pathological compensation from a biologically plausible ion channel expression model. *Neuron* 82(4): E2645–E2654.

O'Leary, T. and D. J. A. Wyllie (2011) Neuronal homeostasis: time for a change? *J Physiol* 589.20:4811–4826

Rose, S. (2012). The need for a critical neuroscience. In *Critical Neuroscience: A Handbook of the Social and Cultural Contexts of Neuroscience,* S. Choudhury and J. Slaby (eds.) Hoboken, NJ: Wiley-Blackwell.

Sporns, O. (2010). *Networks of the Brain.* Cambridge, MA: MIT Press.

Sterling, P. and M. Freed (2007). How robust is a neural circuit? *Visual Neuroscience*, **24**, 563–571.

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology & Behavior* 106:5–15

Sterling, P. and S. B. Laughlin (2015). *Principles of Neural Design*. Cambridge, MA: MIT Press.

Svoboda, K. (2015). Probing Frontal Cortical Networks during Motor Planning. Abstract, Center for the Neural Basis of Cognition, 10 November 2015. http://www.braininstitute.pitt.edu/event/probing-frontal-cortical-networks-during-motor-planning

Zhang, Y. and S. M. Chase (2015). Recasting brain-machine interface design from a physical control system perspective. *J Comput Neurosci* 39:107–118

Eight Myths about Scientific Realism

ABSTRACT: Selective realist projects have made significant improvements over the last two decades. Judging by the literature, however, antirealist quarters seem little impressed with the results. Section I considers the selectivist case and its perceived shortcomings. One shortcoming is that selectivist offerings are nuanced in ways that deprive them of features that—according to many—cannot be absent from any realism "worth having". Section II (the main part of the paper) considers eight features widely required of realist positions, none of them honored by selectivist projects. Modulo those requirements, even if selectivists managed to clear other shortcomings of their project selectivism would still not be a position worth considering. Next the historical background and present credentials of the requirements in question are examined. All are found to rest on myths and confusions about science and knowledge. If this is correct, realists and antirealists should reject the requirements.

I. Background

The antirealist waves of the 1980s stifled naïve realist projects, but they also gave rise to critical realist reactions, particularly a shift in the way theories are accepted at face value from whole constructs to selected "theory-parts" (existence claims, narratives and structures regarding features beyond the reach of unaided perception). Moves in this "selectivist" direction were variously developed in the 1980s and 1990s, most influentially by Leplin (1984), Worrall (1989b), Kitcher (1993), Leplin (1997), and Psillos (1999). Selectivists see in the history of science a past littered not just with failures but also clear successes, especially after the consolidation of methodologies focused on impressive novel prediction in the early 19th century. The successes selectivists point to involve law-like structures all over physics, functional (as opposed to formally "fundamental") entities like the particles invoked by the kinetic theory of matter, numerous extinct species hypothesized by Darwin and his circle, structures and processes from microbiology, much in Mendelian genetics, myriads of molecular structures, and most of the subatomic entities deemed well-established since the 1950s, along countless causal networks, histories and functional entities in virtually all theories with warrant in terms of impressive novel predictive success. Selectivists thus respond to skeptical readings of the history of science with optimistic readings, which they argue are better justified than Laudan (1981)'s skeptical appeals. History, Leplin (1984) noted early in the debate on selectivism, is not opposed to realism any more than our experience of ordinary objects is unambiguously veridical.

In selectivist terms, successful scientific theories may provide imperfect representations of unobservable aspects of some of their intended domains, but they do get those aspects right to some significant extent—and *that is what matters to a realist stance*. Realism has to do with

having warranted augmentative inference at levels that reach into unobservables, i.e. beyond the level allowed by its contrast position—constructive empiricism.

Developing selectivism into a mature project has not proved easy. The initial criteria proposed for identifying theory-parts worthy of realist commitment were either too vague or picked up through "retrospective" projection of current. As Kyle Stanford (2006) cautioned, mere retrospective projection of current science reflects limitations of human imagination as easily as it does truth-content and can be variously misleading; also, it can be self-serving, and worse still it severely weakens selectivism by giving up the traditional realist goal of identifying the truthful parts of a theory while the theory is still alive. Realists need to develop compelling criteria for prospective projection, applicable to theories in full flight, and over the last decade selectivists have moved imaginatively to respond to this challenge. One promising contribution is a stronger emphasis on impressive novel predictions as a marker of success and truth content. This trend is multiply developed in works that revisit in detail the cases most used by antirealists as springboard exemplars of gross epistemic failure, as well as studies of other seemingly germane cases from the last 200 years (e.g. Saatsi 2005, Saatsi & Vickers 2011, Votsis 2011, Vickers 2013, @@@). While the debate is far from over, upgraded proposals are on view in the selectivist analyses just cited. At the very least, the initial antirealist arguments from radical underdetermination and so-called "skeptical inductions" have been weakened by selectivist challenges to the antirealist arguments at work. Still, many critics join Stanford in thinking that selectivism lacks a convincing realist criterion for prospective identification of theory-parts. As said, promising selectivist developments seem on view in this regard, but there is something else.

Something seems to be making the selectivist project intellectually unattractive in some quarters, independently of the issue about the criterion for theory-parts. There is, in particular, a perception (not least among many sympathizers of realism) that selectivism advances its case at the cost of diluting its realist import, resulting in *a stance "not worth having*." By the lights of selective realism, an empirically successful theory T contributes significant truths about unobservables but

(1) typically, what makes T approximately true is that abstract versions of some of its parts are truthful, making the realist stance applicable to selected fragments of T rather than the integral whole initially intended;

(2) such truth as T contains need not have universal applicability;

(3) T need not offer literal truth at its most fundamental level;
(4) the significance of T's central terms is high in unificationist rather than epistemic terms;
(5) T adds significantly to our knowledge of unobservables in the intended domain, but there is no reason to expect T to be "right for the most part" at *any* level (what matters is that it yields epistemic gain at theoretical levels).
(6) T may not instantiate uniformly convergent progress towards any "final description;"
(7) the intelligibility T confers to its intended domain is generally incomplete.

Each of the above tenets clashes head on with widespread assumptions and expectations regarding a realist stance about theories. The latter, many believe, *should* (1) constitute integral wholes, (2) apply universally, (3) give correct theoretical description, (4) have central terms that refer, (5) be, at least, *right for the most part.* (6) display epistemic progress, and (7) offer substantial intelligibility of the intended domains, Behind these expectations about scientific theories and what theoretical claims amount to is a view on what a *realist position worth having* comprises: to be worth having, a realist position must encompass strong versions of most of the listed assumptions. Antirealists (and not a few realists) routinely take these assumptions for granted. This aspect of the debate needs discussion because, as noted, the assumptions in question are clearly at odds with the selectivist strategy, which—generalizing Worrall (2016) a bit—might be *the only viable realist game in town.*

II. Taxing Assumptions

There is a view, shared by numerous scientists, according to which scientific realism cannot be a position worth having unless it encompasses most of the traits listed at the end of the last section. One problem with those traits is that they provide antirealists with fodder for criticizing positions that embrace them and realists for dismissing positions that lack them. Let us consider the listed items in detail.

(1) **Theories as Integral Wholes**. Selectivism rejects the view that theories and conceptual networks are intellectual constructs made of non-separable parts. The integral wholes vision commits realism to nothing less than complete theories. Motivations for it come from at least two

fronts. One includes linguistic holism and/or the statement view of theories, endorsed in the 1960s and 1970s by thinkers as superficially different as Ernest Nagel and Thomas Kuhn. Another motivation, good for a weaker version of the vision, has been the presumption that some concepts are grounded in "metaphysical necessities," a position widely held in natural science until the early 1900s. In the 19th century it was thought that breaking of a theory into independently assertible parts had drastic limits. A case in point was the need felt for positing an ether of light, as at the time waves were conceived of within a traditional metaphysics that regarded them as propagating disturbances and thus as ontologically dependent entities that *required* the existence of something being disturbed (@@@). Institutional deference towards similarly presumed conceptual necessities is massively lower now. One major inflection point was the acceptance of Einstein's Special Relativity, which opened the road to changes in both the conception of light and the requirements of intelligibility in physics.

Nobody thinks now that light is completely as Fresnel or Maxwell imagined, yet— having no conceptual links closed to the possibility of scientific revision— there is little question that Fresnel's theory got many things right, e.g. what might be termed "Fresnel's Core": light is made of microscopic transversal undulations, and these undulations follow the Fresnel laws of reflection and refraction. Abstracted from reference to the wave substratum, this schematic part of the theory spells out a descriptive core that all subsequent theories of light have retained. Once conceptual networks are recognized as relations sustained by revisable inductive conjunctions, scientific "good sense" allows shifts in science towards theory-parts cut out from the rest. There is a historical supplement to this. There has never been much serious allegiance to theory "unbreakability" *in scientific practice*. As scientists developed their ideas, virtually all took a realist stance towards just selected parts of a theory at hand while taking a non-realist stance towards other parts (e.g. Newton's approach towards Kepler's cosmology and Galileo's mechanics; 19th century wave theorists towards particle theories of light, Einstein towards Fresnel's Core, Einstein towards Newtonian mechanics, molecular geneticists towards Mendelian genetics, and so forth). Being selective about what to take at face value in a theory is exactly what selective realists do, also what we all do in ordinary life. The idea that proper theories are unbreakable integral wholes just rests on myth.

(2) **Universality**. Another widespread assumption is that, for realism, proper scientific theories must hold universally. We find this view expressed in e.g. van Fraassen (1980: 86): from a realist perspective, he claims, ‘‘a theory cannot be true unless it can be *extended* consistently, without correction, to all of nature’’

This request rests on myth. There is no reason to think that interesting theories can be so extended even at the lowest phenomenal level. Generalizations limited to the observable level typically turn out to be true only over restricted ranges, just as with theoretical generalizations. The standards of acceptability should not be arbitrarily raised against scientific theories. So, past successful theories could not be extended consistently, without correction, to all of nature. However, as selectivists show, those theories made significant cognitive gains at significant levels, where various assortments of the theoretical descriptions they licensed remain both accurate and illuminating. The universality objection, it seems, burdens realism with a suicidal demand.

(3) **Truthful description**. Realists are allegedly claim that what a theory T says about entities, properties, relations and processes should be construed literally; and to take a realist stance towards T is to believe that what it says is literally true. This view comprises three major lines: (3a) literalism, (3b) accuracy realism, and (3c) a methodological supplement.

(3a) Like their biblical counterparts, theory-literalists think one mistake in a narrative is one mistake too many. Phlogiston theory got some of its central claims wrong, as did also Fresnel’s theory, Mendel theory, Bohr’s 1913 theory of the hydrogen atom, and countless other theories, so those theories were all completely wrong.

The antirealist uses of literalism are straightforward. If departures from literal accuracy, however small, make theories count as different, then the chances of a scientist ever picking *the* right theory will be wretchedly small (argument of the bad lots). And the probability of conjecturing the one (and only one) truthful theory will be hopelessly small (problem of the base rate). And, so, at any given time, the chances that the one truthful theory is among the as yet “conceived alternatives” will be overwhelmingly low.

Happily for realists, the expectations in (3a) belong in fairy-tales. Scientific theorizing is rarely strictly literalist. Scientists effectively abandoned literalism early in modern times, as they

began to articulate explanatory idealizations that carried an expectation that nothing in nature exactly realized them. For example, the aim of the kinetic theory of matter developed around 1860 was to causally account for approximate empirical laws that had been gathered in the two previous centuries about the macroscopic behavior of gasses (e.g. PV = nRT) and materials (e.g. thermal expansion). Crucially, in the case of gases, the accounts invoked structureless point-particles—the so-called "ideal gas"—that the theorists involved did not believe existed in nature. The ideal gas was *explicitly* an idealization, with a two-fold expectation at work: (i) actual gasses are made of non-ideal corpuscles moving at random and located at relatively large distances from one another "on average"; and (ii) the behavior of those actual corpuscles *instantiated that of the ideal gas to a significant degree* within a certain restricted domain. There was no question that ideal gasses literally construed had to be "real" in order to take the theory realistically. Scientific theories are likewise *generally* false in strictly literal fashion. As with maps, the point of realist interest is the extent to which a theory's depictions match the *intended* domain. Theoretical representations of empirical domains resemble maps far more than they do assertions (e.g. Giere 2006). Selectivists proceed accordingly: taking a realist stance towards a theory T amounts to claiming only that some of the explanations and descriptions distinct to T are correct by *acceptable standards*.

(3b) In mathematized disciplines literalism easily ups its ante. According to a long lived assumption of quantitative exactitude, there are in nature quantities of which concrete systems have definite values, and in a correct theory the claims it makes correspond to the world with total accuracy. This ideal is found in early modern scientists, notably theorists with strong Platonist leanings such as Kepler.

Dear though these expectations of divine accuracy and depth are, they rest on myth. Such correspondence as mathematized theories have to the world is not conditioned to radical accuracy. As Bertrand Russell noted on behalf of sound epistemology,

> "Although this may seem a paradox, all exact science is dominated by the idea of approximation. When a man tells you that he knows the exact truth about anything, you are safe in inferring that he is an inexact man. Every careful measurement in science is

> always given with the probable error [...] every observer admits that he is likely wrong, and knows about how much wrong he is likely to be." (1931: 42)

More recently, in a more comprehensive vein, Paul Teller (2015) complains that "accuracy realism" assumes that the quantities invoked by a theory actually refer. But—he notes—this misunderstands the fabric of theoretical representation, because theories generally formulate *idealizations* that burden quantitative attributions with failure of specificity in picking concrete cases. In the narrowest literal sense, the claim "the meter-standard kept in Paris is 1 meter long" may be true only by *definition*—any attempt to check it with absolute precision against any external objective length would be frustrated by, to begin with, ineliminable thermal and quantum mechanical fluctuations. The point is that one-to-one matching makes no sense as a goal in scientific language, given that so many descriptive words in science are intrinsically vague and/or refer to idealizations. Actual reference to lengths presumes just perspectivally *acceptable* (never absolute) accuracy. At the lowest empirical levels also, completely exact assertions are generally neither relevant nor true. This connects with a related point, namely, the *irrelevancy* of these literalist and accuracy assumptions to the actual realism/antirealism debate. Shaped by the discussions started in the 1980s, the dispute is now primarily about whether or not warranted augmentative scientific inferences reach into unobservable domains. Ordinary realism about chairs, cats and mountains fails the ideals of radical literalism and accuracy no less than scientific realism.

(3c) The methodological supplement claims that science would be merely an instrumentalist affair unless theorists aim to produce a complete description of the way things are, with scientists as pursuers of God-like reportage (perfect "mirror reflection"): scientific theories advance towards the truth, all the truth, and nothing but the truth (see e.g. Sankey 2008's discussion of this). Although this position lost much of its ancient appeal in the 18th century, to this day some top theoreticians continue to wax lyrical expressing it, especially in "editorials".

> "The 'theory of everything' is one of the most cherished dreams of science. If it is ever discovered, it will describe the workings of the universe at the most fundamental level and thus encompass our entire understanding of nature. It would also answer such

> enduring puzzles as what dark matter is, the reason time flows in only one direction and how gravity works. Small wonder that Stephen Hawking famously said that such a theory would be 'the ultimate triumph of human reason – for then we should know the mind of God' ". (New Scientists, 4 March 2010[1])

This colorful supplement lacks warrant if, as selectivists claim, the realist stance can be consistently and fruitfully applied to selected theory-parts.

The realist badge of honor is not awarded for telling the truth, all the truth, and nothing like the truth about anything—let alone reading the mind of God. It is a distinction for *finite cognitive achievements forged with crooked tools*. See also (6) below.

(4). **Realist Significance of the "Central Tenets" of a Theory**. A related common assumption is this: Even if truthful description may have limits, taking a theory T realistically requires commitment to T's central tenets (i.e. those about the entities, principles and laws that individuate T). In Laudan's version, realism about T commits to the view that the T's central terms *successfully refer*.

There is little question that in numerous scientific theories the central terms fail to refer—on this point we all have a debt of gratitude to Laudan. However, once theories are no longer approached as unbreakable wholes the emphasis on central terms wanes. If anything, the reference that matters is that of theory-parts. Then, on the explanatory side, the scientific focus is on the structures of possibilities of its intended domain D. As such, a theory is not exclusively about the entities and relations invoked at the level of its central terms. Primarily the theory is about D, whose relevant entities and structures include those that may be found at intermediate levels of description—like Fresnel's Core. A theory may thus be individuated by its central tenets, but the latter do not exhaust the theory's realist import. The appropriate realist focus is those theoretical claims derivable from the theory and for which there is strong evidence (and so a strong expectation of truthfulness), not whether the terms involved are "central", "intermediate", or "peripheral".

---

[1] "Knowing the mind of God: Seven theories of everything", New Scientists, 4 March 2010: https://www.newscientist.com/article/dn18612-knowing-the-mind-of-god-seven-theories-of-everything/

(5). **Being "right for the most part".** Another related assumption links the realist stance towards a theory with the claim that the theory is right "for the most part". Michael Devitt, for example, voices this assumption when he defines scientific realism as the doctrine according to which "Most of the essential unobservables of well-established current scientific theories exist mind-independently and mostly have the properties attributed to them by science" (2005: 769). In his view, theories that are well-established theories *by today's* methodological standards are right *for the most part*.

This supposition sounds reasonable at first hearing but it too seems suicidal for realism. Virtually all the past theories realists want to be realist about seem to have turned out to be wrong "for the most part"—unless "being right" is granted with postmodern largesse. Newtonian mechanics is "right" for a comparatively tiny regime of speeds and fields. Bohr's theory of the atom gets impressive aspects right but otherwise is wrong for the most part of the entire quantum domain. Mendel's theory invites a similar reaction. For all we know, our excellent present physics may be wrong for most of the *total universe*. So, scientifically successful theories seem "wrong for the most part". But they have great realist import, nonetheless. That import comes from the fact that they get right novel *significant unobservable* aspects of their intended domains. As David Bohm urged long ago, piecemeal caution needs to be exercised in one's realist commitment to the entities, regularities and processes invoked by well-established current scientific theories (1957, Chapter V). Two lines of reasoning in particular support this prudence (@@@): (1) Qualities, properties of matter, and categories of laws expressed in terms of some finite set of qualities and laws are generally applicable only within limited contexts (in terms of ranges of conditions and degrees of approximation). (2) There is no reason to suppose that new qualities and laws will *always* lead to mere correction refinements that converge in some simple and uniform way. This may occur in some contexts and within some definite range of conditions, but in different contexts and under changed conditions the qualities, properties and laws may be quite novel and lead to dramatic effects relative to what previous theorizing would have led to expect. For example, for bodies moving with speeds negligible compared to the speed of light, the laws of relativity lead to small corrections of the laws of Newtonian mechanics. But they also lead to such qualitatively new results as the "rest energy" of matter. Further laws yet to discover may be vastly more bizarre.

(6) **Progress**: The realist expectation that successful science achieves cumulative truth content about unobservables is frequently nailed to the idea that "modern science is converging on a single picture of the world". Claims along these lines come in several flavors, in particular (a) linear epistemic progressivism and (b) "metaphysical" realism.

(6a) Convergent progress. Léo Errera expressed the idea in his *Botanique Générale* of 1908: "Truth is on a curve whose asymptote our spirit follows eternally[2]." This expectation has recurrent mystical roots in science. John Herschel, for example, is cited by Marcel de Serres as saying "All human discoveries seem to be made only for the purpose of confirming more strongly the truths come from on high, and contained in the sacred writings[3]."

Convergent progressivism runs against a recurrent realization in modern science. As selectivists recognize, successful theories give knowledge but they usually err at numerous levels of description. Successful theories don't give us everything there is to know about any intended domain, let alone 'The World.' Finite sets of simple laws can provide correct descriptions and predictions when we constrain their context enough, notes Bohm (1957), but we should expect unrestricted theories to be false. Many defenders of scientific objectivity have followed suit, stressing the shift from traditional searches for a comprehensive world-view to explicitly perspectival searches for piece-meal knowledge about domains of current scientific interest, leading to assertions of corresponding partiality.

(6b) In no better shape is the claim that realism is committed to the existence of one true and complete description of the world, whose truth bears one-to-one correspondence to 'mind-independent reality, so that the purpose of science is to discover that description. Critics persuasively dismiss this brand or realism. But no knowledgeable realist has held such a position in generations. It is a thesis recalled from the grave in the late 1970s and 1980s by Hilary Putnam under the label "metaphysical realism," a view he presented as an example of a hopelessly jumbled project (e.g. Putnam, 1978: 49, and 1990: Preface).

---

[2] *Recueil d'Œuvres de Léo Errera: Botanique Générale* (1908), 193. As translated in John Arthur Thomson, *Introduction to Science* (1911): 57

[3] Marcel de Serres, 1845. "On the Physical Facts in the Bible Compared with the Discoveries of the Modern Sciences". *The Edinburgh New Philosophical Journal* (Vol. 38): 260. [239-271]

(7) **Intelligibility**: Another claim often associated with realism is that science aims to provide truthful explanations that make the phenomena at hand intelligible. This condition comes in (a) radical and (b) moderate strengths. The radical version calls for explanations that leave the intellect content and with no further whys. The weak condition calls for explanations that make the target phenomena *more* but not necessarily fully intelligible.

(7a) Leibniz's rationalist objection to Newton's Theory of Gravitation exemplifies the radical version. He complained that if gravity were thought as a real force, then its effect would be a mysterious action at a distance. Leibniz blamed Newton for introducing "occult" forces into science, and until the end of his life Newton hoped to produce a properly "intelligible" account of gravity involving only action by contact interactions—he did not succeed. Modern scientific theories do not provide radical intelligibility. Once Galileo gave up his initial hope of presenting inertial motion as uniform circular motion, the theory of free fall he accepted left open at least as many whys as it closed. Why or how Galileo's mysterious mathematical structures arise in nature? The same goes for subsequent theorizing. Why or how the regularity given as Newton's law of gravitation arises? Why or how Fresnel's Core arise? Why or how the speed of light is a universal invariant? Contemporary fundamental theories fail radical intelligibility just as clearly.

Realists need not worry about this. Calls for radical intelligibility rest on views of cognition now widely recognized as mythical. Barring mystical insight and such, all actual understanding comes with opaque spots. At every scientific stage scientific warrant (and intelligibility) stops somewhere, albeit usually not at the traditional empiricist boundaries. Realism is compatible with suspending judgment about whether a certain theoretical claim correctly describes a fundamental or derivative aspect of nature. This is exemplified in the stance realists take towards e.g. Fresnel's Core, the invariance of light's speed, and fundamental principles in general.

A theory that saves all the known phenomena but whose reliable parts comprise only structures and explanations at phenomenal levels, provides the lowest level of understanding. This makes for a constructive empiricist take, which escapes skepticism by accepting realism about just the theory's empirical substructures. The point here is that radical theoretical intelligibility is not necessary for taking a realist stance towards a theory. From a selectivist

perspective, the key factor for taking a theory-part realistically is not the "intelligibility" it confers but its indispensability for maintaining the theory's predictive power in the context of current *background knowledge.* Ptolemaic orbits were denied realist interpretation not primarily because they failed the intelligibility requirement—Ptolemaic constructions went out of their way to honor, of all requirements, *intelligibility* (then guided by the Principle of Uniform Circular Motion for heavenly bodies and the Aristotelian arguments for the fixity of the Earth). Rather, Ptolemaic orbits were refused realist interpretation because the epicycles, deferents and equants they invoked were grossly *underdetermined by extant knowledge* (i.e. available data and cosmological principles). Positive evidence for the orbits specifically proposed was lacking.

None of this is not to question the realist relevance of theories that seek to achieve deep understanding. What is denied is that *scientific* realism must embrace radical intelligibility. Radical intelligibility is a trait realism about observables and every day affairs neither honors nor is expected to honor.

(7b) This brings us to cogent versions of the moderate intelligibility condition. Selectivists take a realist stance only towards theory-parts deemed to be both indispensable for the theory's success and free of compelling specific doubts against them (@@@). That is, the realist stance goes *only* to tenets for which there is strong positive evidence by modern scientific standards. In all the cases highlighted by realists, the selections supported by the strongest level of evidence available make the target domain intelligible well beyond the observable levels. When, by contrast, the positive evidence for a theory does not reach the unobservable explanatory posits that make the relevant phenomena intelligible, then the best stance to take about the theory is not realism but *constructive empiricism*. This clarifies what introductory characterizations of scientific realism get right about the intelligibility condition: A good theory must not have just significant predictive power but must also make the relevant phenomena *intelligible* (Richard DeWitt 2010: 72). If the theory parts that do this lack evidential warrant, then the reasonable stance towards them is constructive empiricism.

(8) **Realism Worth having.** Topping the above assumptions, there is a popular notion to the effect that a realist stance failing to adhere to most of the above requirements is "*not a realism worth having*". Against this idea, I have argued that none of the listed assumptions is worth

having. Every one of them lacks convincing warrant. Moreover, even if the assumptions did get proper warrant they face a deeper problem: the assumptions are *irrelevant* to the current realism/antirealism debate—they do not expose relevant contrasts between inferences limited to the phenomenal level and inferences that reach into theoretical levels.

In modern science, virtually all interesting augmentative inferences violate the listed assumptions. So, the latter simply and arbitrarily raise the epistemological standards of acceptability against theoretical assertions. If the above considerations are correct, then, realists and antirealists should reject the assumptions examined in this paper—they all rest on counterproductive myths and confusions.

References

Bohm, David (1957). *Causality and Chance in Modern Physics*. London: Routledge & Kegan Paul Ltd.

Devitt, Michael (2005). "Scientific Realism". In *The Oxford Handbook of Contemporary Philosophy*, Frank Jackson and Michael Smith, eds. Oxford: Oxford University Press: 767-91.

Giere, Ronald N. (2006). *Scientific Perspectivism.* Chicago: University of Chicago Press.

DeWitt, Richard (2010): *Worldviews*. Malden, MA: Wiley-Blackwell.

Kitcher, Philip, 1993. The Advancement of Science. Oxford: Oxford University Press.

Laudan, Larry. 1981. "A Confutation of Convergent Realism". *Philosophy of Science* 48: 19-49.

___. 1984. *Science and Values*. Berkeley: University of California Press.

___. 1996, *Beyond Positivism and Relativism: Theory, Method and Evidence*. Boulder, CO: Westview Press.

Leplin, Jarrett, ed. 1984. *Scientific Realism.* Berkeley: University of California Press.

___. 1997. *A Novel Defense of Scientific Realism.* New York: Oxford University Press.

Psillos, Stathis. 1999. *Scientific Realism: How Science Tracks Truth.* New York: Routledge.

Putnam, Hilary (1978).

Russell, Bertrand, 1931. *The Scientific Outlook*. London: George Allen & Unwin, Ltd.

Saatsi, Juha (2005). "Reconsidering the Fresnel-Maxwell Case Study." *Studies in History and Philosophy of Science* 36 (3): 509–38.

Saatsi, Juha and Peter Vickers (2011). "Miraculous Success? Inconsistency and Untruth in Kirchhoff's Diffraction Theory." *British Journal for the Philosophy of Science* 62: 29–46.

Stanford, P. Kyle. 2006. *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford: Oxford University Press.

Teller, Paul (2015). "Language and the Complexity of the World;" forthcoming.

Van Fraassen (1980). The Scientific Image. Oxford: Oxford University Press.

Vickers, Peter (2013). "A Confrontation of Convergent Realism". *Philosophy of Science* 80: 189-211.

Votsis, Ioannis (2011). "Saving the Intuitions: Polylithic Reference." Synthese 180 (2): 121–37.

Worrall, J. 1989a. "Fix it and be Damned: A Reply to Laudan". *British Journal for the Philosophy of Science* (40): 376-388.

------ (1989b). "Structural Realism: The Best of Both Worlds". *Dialectica* 43: 99-124.

------ (2016). "Structural Realism – the Only Viable Realist Game in Town." Forthcoming in *Scientific Realism: Objectivity and Truth in Science,* Wenceslao Gonzalez & Evandro Agazzi, (eds.).

Concrete Models and Holistic Modelling*

Wei Fang[ϕ]

Department of Philosophy, University of Sydney

Abstract: This paper proposes a holistic approach to the model-world relationship, suggesting that the model-world relationship be viewed as an *overall structural fit* where one organized whole (the model) fits another organized whole (the target). This approach is largely motivated by the implausibility of Michael Weisberg's weighted feature-matching account of the model-world relationship, where a set-theoretic conception of the structures of models is assumed. To show the failure of Weisberg's account and the plausibility of my approach, a concrete model, i.e. the San Francisco Bay model, is discussed.

* Draft paper, please do not quote without permission.

ϕ Address: University of Sydney, NSW 2006, Australia. Email: wfan6702@uni.sydney.edu.au.

1. Introduction

One philosophical interest in the philosophy of modelling focuses on the problem of the model-world relationship, also known as the representation problem. Among many approaches to this problem, the similarity account has attracted much attention recently. Ronald Giere (1988, 1999a, 1999b, 2004, 2010), Peter Godfrey-Smith (2006) and Michael Weisberg (2012, 2013) have made the most substantial contributions.

The core of this account, first developed by Giere, is a view of the model-world relationship:

> The appropriate relationship, I suggest, is *similarity*. Hypotheses, then, claim a *similarity* between models and real systems. But since anything is similar to anything else in some respects and to some degree, claims of similarity are vacuous without at least an implicit specification of relevant *respects and degrees*. The general form of a theoretical hypothesis is thus: Such-and-such identifiable real system is similar to a designated model in indicated respects and degrees. (Giere 1988, 81; author's emphasis)

However, critics point out that this account is only schematic since it falls short of specifying the relevant *respects* and *degrees* (Suárez 2003). Moreover, Giere argues that a philosophical account of scientific representation should also take into consideration factors such as the *roles* played by scientists, and the *intentions* those scientists have when modelling (Giere 2004, 2010). Given these considerations, Weisberg develops a more sophisticated similarity account, called the *weighted feature-matching* account

(2012, 2013). The basic idea of his account comes from psychologist Amos Tversky's *contrast* account of similarity, which states that the similarity of objects *a* and *b* depends on the features they share and the features they do not. In light of this, Weisberg proposes his own account:

S (m, t) =

$$\frac{\theta f(M_a \cap T_a)+\rho f(M_m \cap T_m)}{\theta f(M_a \cap T_a)+\rho f(M_m \cap T_m)+ \alpha f(M_a - T_a)+ \beta f(M_m - T_m)+ \gamma f(T_a - M_a)+ \delta f(T_m - M_m)} \quad (1)$$

f(x) refers to the weighting function, α, β, γ, δ, θ, and ρ denote weighting terms (parameters), subscripts *a* and *m* stand for attributes and mechanisms,[1] and *M* denotes the model and *T* the target. ($M_a \cap T_a$) stands for attributes shared by the model and the target, ($M_a - T_a$) attributes that the model has while the target does not, and ($T_a - M_a$) attributes that the target has while the model does not. The same story goes for mechanisms *m*. Attributes and mechanisms as a whole are called *features* of the model and the target.

An interpretation for this equation is needed. First, there must be a feature set Δ, and the set of features of the model and the set of features of the target are defined as sets of features in Δ. The elements of Δ are determined by a combination of context, conceptualization of the target, and the theoretical goals of the scientist. Besides, the

[1] Properties and patterns of systems are termed attributes, and the underlying mechanisms generating these properties and patterns are termed mechanisms (Weisberg 2013, 145).

contents of Δ may change through time as science develops, which in turn might result in a reevaluation of the established model-world relationship (*Ibid.*, 149).

Second, consider the values of weighting parameters α, β, γ, δ, θ, and ρ. On Weisberg's account, different kinds of modelling require different weighting parameters. For example, if what interests us is the *minimal modelling* which concerns merely the mechanism responsible for bringing about the phenomenon of interest, the goal of this modelling is written as:[2]

$$\frac{|M_m \cap T_m|}{|M_m \cap T_m| + |M_a - T_a| + |M_m - T_m|} \rightarrow 1 \qquad (2)$$

Finally, consider the weighting function $f(x)$, telling us the relative importance of each feature in the set Δ. Weisberg says scientists in most cases have in their mind some subset of the features in Δ, which they regard as especially important. Hence some features are weighted more heavily, while others are equally weighted. Besides, the background theory determines which features in Δ should be weighted more heavily. If the background theory is not rich enough, deciding which should be weighted more heavily is partly an empirical problem.

Having presented an outline of Weisberg's account, I will now argue that this account fails to capture the relationship between concrete models and their targets. To illustrate this

[2] Weisberg also describes three other kinds of modelling requiring different weighting parameters: hyperaccurate, how-possibly and mechanistic modelling (2013, 150-52).

shortcoming (Sec. 3), I will first describe the San Francisco Bay model (Sec. 2). Sec. 4 will propose a holistic alternative to Weisberg's account, suggesting that the model-world relationship be viewed as an *overall structural fit* where one organized whole fits another organized whole. Sec. 5 will examine a case where the organization of the whole can be treated as simply another feature.

2. The San Francisco Bay Model

John Reber worried about the fragility of the water supply in the San Francisco Bay area in the 1950s. To solve this problem, he proposed an ambitious proposal, namely, to dam up the Bay. Carrying out this plan would not only supply San Francisco with unlimited drinking water but also entirely change the area's transportation, industrial, military and recreation landscape (Weisberg 2013, 1). However, his critics worried that Reber's plan would only achieve its aims at the cost of destroying commercial fisheries, rendering the South Bay a brackish cesspool, creating problems for the ports of Oakland, Stockton, and Sacramento, and so on (Jackson and Peterson 1977; Cf. Weisberg 2013, 1).

To settle this dispute, the Army Corps of Engineers was charged with investigating the overall influence of the Reber plan by building a massive hydraulic scale model of the Bay (Weisberg 2013, 1-2). Once the model was built, it was adjusted to accurately reproduce several measurements of the parameters such as tide, salinity, and velocities actually recorded in the Bay (for details see Army Corps of Engineers 1963). After the adjustment, it was time to verify the model:

> Agreement between model and prototype for the verification survey of 21-22 September 1956, and for other field surveys, was excellent. Tidal elevations, ranges and phases observed in the prototype were accurately reproduced in the model. Good reproduction of current velocities in the vertical, as well as in the cross section, was obtained at each of the 11 control stations in deep water and at 85 supplementary stations. The salinity verification tests for the verification survey demonstrated that for a fresh-water inflow into the Bay system […], fluctuation of salinity with tidal action at the control points in the model was in agreement with the prototype (Huggins and Schultz 1967, 11).

After the verification, modellers were in a good position to assess the Reber plan through the model built. The investigation showed that it would considerably reduce water-surface areas, reduce the velocities of currents in most of South San Francisco Bay, reduce the tidal discharge through the Golden Gate during the tidal cycle, and so forth (Huggins and Schultz 1973, 19). Given these disastrous consequences, the Army Corps then denounced Reber's plan (Weisberg 2013, 9).

3. How Could Weisberg's Account Shed Light on the Bay Model?

I have argued elsewhere that Weisberg's account cannot shed light on mathematical models due to its atomistic conception of features and its assumption of the set-theoretic approach to model structures (citation anonymized). I find that the same charges can be raised in the case of concrete models.

Consider the first charge: Weisberg's account is committed to an atomic conception of features. The key of Weisberg's account is the claim that the similarity of objects *a* and *b* depends on the features they share and the features they do not share. Let us take a closer look at the equation (1). The numerator invites us to weight features shared, and the denominator asks us to weight all features involved (including three feature subsets: features shared, features possessed by the model but not the target, and features possessed by the target but not the model). Each feature is weighted independently and only once, with it falling into one of the three feature subsets. The numerator is the weighted sum of features shared, the denominator is the weighted sum of features shared and unshared, and the similarity measure is the ratio of the numerator to the denominator.

However, features in the Bay model are not atomistic and independent of each other. As Huggins and Schultz put it explicitly, "Among the problems to be considered were the conservation of water […]; […] the tides, currents and salinity of the Bay as they affect other problems […]. None of these problems can be studied separately, for each affects the others" (1973, 12). The reason why none of these problems can be studied separately is because factors involved in these problems cannot be studied separately.

Consider, for instance, the relationship between two key features in the model: tide and salinity. Salinity levels vary along an estuary depending on the mixing of freshwater and saltwater at a site. An estuary "is the transition between a river and a sea. There are two main drivers: the river that discharges fresh water into the estuary and the sea that fills the estuary with salty water, on the rhythm of the tide" (Savenije 2005, Preface ix).

To illustrate this "rhythm of the tide", consider the effect of the spring-neap tidal cycle on the vertical salinity structure of the James, York and Rappahannock Rivers, Virginia, U.S.A.:

> Analysis of salinity data from the lower York and Rappahannock Rivers (Virginia, U.S.A.) for 1974 revealed that both of these estuaries oscillated between conditions of considerable vertical salinity stratification and homogeneity on a cycle that was closely correlated with the spring-neap tidal cycle, i.e. homogeneity was most highly developed about 4 days after sufficiently high spring tides while stratification was most highly developed during the intervening period. (Haas 1977, 485)

This short report shows not only that characteristics of salinity (such as stratification and homogeneity) are influenced by characteristics of the tide, but also that there is a phase connection (or synchronization) between tidal cycle and salinity oscillations. The former is a causal relationship while the latter is a temporal relationship. The phase connection among features was also emphasized by the Army Corps when verifying the Bay model, saying "These gages were installed in the prototype and placed in operation several months in advance of the date selected to collect the primary tidal current and salinity data required for model verification, since *it was essential to obtain all data simultaneously for a given tide over at least one complete tidal cycle of 24.8 hours*" (1963, 50; my emphasis). Moreover, the same story goes for tide and tidal currents (for details see Army Corps 1963, 20).

In short, features in a model bear not only causal relationships, but also temporal relationships to one another. This implies that, when verifying the model, features of the

model causally interact with each other in producing certain outputs (e.g. predictions, effects, phenomena, etc.), rather than that they individually or separately produce outputs. So although outputs of key features in the Bay model can be identified and measured separately, they are not produced separately.

It is important to note that the causal interaction among features may lead to a different kind of interaction, i.e. a "similarity interaction",[3] wherein features interact with one another in producing the similarity value. That is, one feature's contribution to the similarity value depends on other feature(s)' contribution to that value.[4] The difference between causal and similarity interaction is that the latter is a statistical relationship among measured features, and can be viewed as a reflection of the former when coupled with an assumption that there might be such an underlying causal structure.[5] For example, a similarity interaction is shown by the verification of salinity in the Bay model, where the measurement of salinity (as a measurement of one feature's contribution to the similarity

[3] I thank X for suggesting this term for me.

[4] This point can be best illustrated with the curve fitting example: when computing the fit of a straight line $y=ax+b$ to a cloud of points, a and b will depend on each other to produce the best fit (I thank X for giving me this example).

[5] This assumption is important because there are cases where the fact that there is similarity interaction cannot guarantee that there is also causal interaction, because some randomly generated data set may also show interaction among features. In other words, causal interaction can lead to similarity interaction and the reverse is not true (I thank Y for letting me know this). I will discuss this assumption, called "precondition" later, in Sec. 4.

value from Weisberg's perspective) depended on other features in the way in which other features were kept constant: "salinity phenomena in the model were in agreement with those of the prototype *for similar conditions of tide, ocean salinity, and fresh-water inflow*" (*Ibid*., 54; my emphasis).

The way that similarity interaction reflects causal interaction, when coupled with the assumption mentioned above, can be expressed as follows: if what is under verification is a causal structure to which modellers do not have direct access (so the structure cannot be a feature in Weisberg's formula), then the coherent behavior of features (i.e. their similarity interactions such as phase connections) is a way of verifying, or at least indicating, the causal interactions in the underlying causal structure.[6] That is the reason why it was so essential to obtain all data simultaneously within a complete tidal cycle for the Bay model, and why all other features must be kept constant when verifying salinity (or other features).

Given features' causal interactions in the model and their similarity interactions when measuring them, it seems that assessing the relationship between a model and its target cannot be simply achieved in the way suggested by Weisberg's equation, for features' contribution to the similarity relationship is not *additive* but *interactive*. That is, to assess the relationship between a model and its target, one cannot measure each feature's contribution independently and then add them together.

4. Set-Theoretic or Non-Set-Theoretic? A Holistic Alternative

[6] I thank X for bringing this point to my attention.

Now we arrive at the problem of why Weisberg's account is deeply committed to an atomistic conception of features. As I have argued elsewhere, this problem ultimately comes down to Weisberg's understanding of the structure of models (citation anonymized). Weisberg says models are *interpreted structures* (2013, 15), so concrete models are interpreted concrete structures. At first glance, I have no quarrel with this understanding. On closer inspection, however, it can be shown that Weisberg's account on the model-world relationship assumes a set-theoretic approach to the structure of models.[7] This is because Weisberg's similarity measure can be derived from the *Jaccard similarity coefficient* between two sets, a coefficient assuming a set-theoretic conception of objects (citation anonymized).

The key to the set-theoretic approach to structures is its assumption that elements of objects (i.e. models and targets) are independent of each other, just as elements of a set are independent of each other. In other words, it construes both the model and the target as a set of independent elements, the similarity between which consists in the ratio of the number of elements shared to the number of all elements (citation anonymized). However, as discussed in Sec. 3, features are not independent. More importantly, their causal interactions may result in a similarity interaction among features.

This similarity interaction undermines Weisberg's account, for it cannot properly capture the dependence relationship of features' contribution to the overall similarity

[7] Note that Weisberg *explicitly* objects to the set-theoretic approach to models (2013, 137-42). However, I think it is compatible to claim that someone *implicitly* assumes what someone explicitly rejects.

measure between a model and a target. Nonetheless, there is still a way to save the very intuitive notion of similarity, by abandoning the set-theoretic conception of structures. That is, if the structure of a model is viewed as an *organized whole* in which each component of the whole is interconnected to other component(s) (directly or indirectly) in such a way that they interact with one another in producing certain phenomena of interest (i.e. outputs). Under such an understanding, therefore, assessing the relationship between a model and its target cannot be simply achieved by assessing each individual feature's relationship and then adding them together. Nor can this be done by assessing each connection among two or more features and then adding them together, even if connections (causal or non-causal) are also interpreted as features. On the other hand, however, the notion of similarity can be minimally preserved by claiming that assessing the similarity or *fit* (I will use fit hereafter) between a model and a target amounts to assessing the *overall structural fit* between the model and its target.

Generally speaking, structural fit means the structure of the model fits the structure of the target *as an organized whole*. That said, nevertheless, it should be stressed that there is no univocal meaning for the term "structural fit" that could encompass all circumstances, nor can a single equation or formula capture all situations. This is largely due to the heterogeneity of modelling practice and its multifarious goals. On the other hand, however, instructive points can still be asserted. In what follows I will elaborate some basics regarding the conception of "structural fit".

Structural fit in mathematical modelling means different things than in concrete modelling. For example, in a very simple case of curve fitting where a straight line $y=ax+b$

is fitted to a cloud of points, features *a* and *b* will interact with each other to produce the best fit. That is, what fits the cloud of points is the overall structure, not the additive sum of each individual feature. As I have argued elsewhere, in more complicated mathematical modelling such as the *maximum likelihood estimation*, the fit is usually achieved through comparing the predicted data set derived from the model *as a whole* to the observed data set derived from the target system (citation anonymized). Individual features of the model simply disappear, and causally related features, as constituting a whole, that co-occur in the data set are what really matters.

In the case of concrete modelling, admittedly, the claim that assessing the fit between a model and a target amounts to assessing the overall structural fit seems to be less apparent. On closer examination, however, the same claim still holds. Let us go back to the verification of the Bay model. At first glance, it seems the verification of the model was achieved by independently verifying the output (i.e. data sets) of each individual feature, as the report showed (see Sec. 2 for the verification report). That is, it seems that by verifying that each feature in the model fits its counterpart in the target, scientists made the judgment that the model fits the target system.

Underlying this seemingly plausible reasoning, however, there remains the problem of why we are allowed to confirm the verification of the model by means of only verifying several outputs of individual features. Or, to put it slightly differently, in terms of what does the fit of features guarantee the judgment about the fit of the model to the target? I take it that it is more than the fit of individual features themselves that makes sense of the reasoning that the model fits the target. There must be a precondition for this reasoning

(remember the "assumption" made in the last section). After all, there are many cases in which the fit of features does not guarantee the fit of the model itself to the target. For instance, a drawing of Tom's face may accurately capture all features of his face, e.g., nose, eyes, mouth, etc., but still falls short of fitting his face, because of the wrong organization of these features, e.g., putting the mouth in between the eyes and nose (Weisberg would argue that the organization could be a feature. I will discuss this point in Sec. 5.).

So if the fit of features is insufficient to vindicate the fit of a model to its target, what could provide this vindication? My claim is, contrary to Weisberg, that it is the *overall structural fit* of the model to the target system that warrants the fit judgment about the model and its target. In other words, the fit of individual features can only succeed in supporting the fit of the model to the target by the precondition that these features can be organized into the whole (i.e. the assumption that there is such an underlying causal structure), not the other way around.

To understand this "holistic reasoning", let me articulate the specifics involved step by step. We first build a concrete model, i.e. a concrete structure, wherein features are interconnected with one other in such a way that they have the potential to interactively produce certain phenomena of interest (i.e. outputs). Before verifying the model, we need to adjust key features to make sure the model works very well. Note that any adjustment will not simply be the adjustment of individual features but also of their interconnections, resulting in the adjustment of the overall structure of the model. Finally, we verify the model by comparing the outputs of the model to the outputs of the target. As with mathematical models, this verification is also usually made via comparing data sets, as

shown in the Bay model. Note that though these outputs can be identified, derived and measured independently, it is causally connected features that interact in producing them. In other words, although you verify each feature separately, the support provided by a single feature is not confined to that feature of the model, but confirms all aspects of the model that are involved in generating that output.

Thus understood, therefore, the gist of verifying a concrete model such as the Bay model can be captured as follows. The verification of each feature, as a component of a whole, is simply the verification of one aspect of the structure. So the verification of different features is the verification of the same structure from different perspectives. Thus, if the model is an organized whole, then the more features that are independently verified the more likely it is that the model resembles the reality. On the other hand, if what is under verification is not an organized whole but an aggregation of independent items, then the verification of each lends no credence to other parts of the aggregated whole—because these items are not causally linked, the verification of each item is only the verification of that item itself.

In sum, the relationship between a concrete model and its target is a holistic matter wherein an organized whole fits (to a certain degree) or fails to fit another organized whole. Though it seems at first blush that the verification of the whole results from the sum of the verification of each component, the real picture is just the reverse: the whole is always in place and the component can gather force in supporting the verification of the whole only when it can be organized into the whole.

5. Organization and Features

As mentioned above, Weisberg would argue that the organization could be a feature, so a drawing of Tom's face capturing accurately not only his nose, mouth, eyes but also their organization can be a good model of Tom's face. A holistic account agrees that organization could be a feature, but disagrees with the way that organization is treated in Weisberg's similarity measure. Intuitively, we may say that a drawing of one person's face is a good model if it has the right features: such as a nose, a mouth, eyes, and the organization of all of these. So it seems that if you get each individual feature right, then you get the whole model right. That is, features *additively* contribute to the goodness of the model.

This intuitive way of understanding scientific modelling, however, obscures the fact that features may interact in producing the fit of a model, as shown in Sec. 4. To reiterate this point and to draw a connection to our current discussion, consider another ordinary example.[8] Suppose Anne's face is an ideal one which scientists want to model. Anne has an ideal nose, which is straight, in contrast to a non-ideal nose, which might be bumped or concave. She also has an ideal nostril, which is round, in contrast to a non-ideal one, which might be triangular or square. Scientist A draws a face for Anne that has a round nostril and a concave nose, while scientist B draws a face that has a triangular nostril and a bumped nose. Drawing A has an ideal feature (the round nostril), but neither feature of drawing B is ideal. Now we ask which drawing better fits Anne's face. It is likely that we

[8] I thank X for giving me this nice example.

will say that B is better because our contemporaries' taste tells us that there is no face so ugly as one with a round nostril and a concave nose, though a round nostril itself is ideal. Hence we see a case wherein the nostril and nose interact to produce the fit of a model to a target.

This discussion leads to a more general question: what are features? In Weisberg's account, a model can *more or less* fit a target, but features are either shared or not. Yet as Wendy Parker points out, "relevant similarities often seem to occur at the level of individual features, not just at the level of the model" (2015, 273). This is because features themselves can be objects such that they more or less fit each other.[9] Weisberg may argue that this problem can be fixed by the assumption that a feature can be redescribed as a set of sub-features, so the similarity between two features can be measured as the result of the similarity between their sub-features. However, I see this treatment as a non-starter, for the similarity between sub-features may also be a matter of degree such that it should be measured as the result of the similarity between their sub-sub-features, and between their sub-sub-sub-features, and so on.

On the other hand, a holistic account does not encounter this problem: if a feature is an object, then it can be viewed as an organized whole. So the relationship between a feature in a model and a feature in a target also consists in their structural fit. Take a minimal model for instance. Most minimal models primarily attempt to represent repeatable patterns of behavior largely insensitive to underlying microscopic details (Batterman 2002, 27). Suppose we are interested in the buckling behavior of struts, and write a

[9] I thank X for bringing this to my attention.

phenomenological formula, called Euler's formula, to characterize it (see Batterman 2002 for details). It seems the pattern of behavior is the only feature involved in this case, i.e., a dependence relationship among several parameters. So assessing the fit between the model and the target comes down to assessing the fit between the feature in the model and the feature in the target. For this, a holistic account can easily come through: the relationship is an overall structural fit, wherein a dependence relationship as a feature fits another dependence relationship.

6. Conclusion

This paper has shown that the assumption of a set-theoretic approach to structures makes Weisberg's account fail to shed light on the San Francisco Bay model. Alternatively, a holistic approach to models, viewing the model-world relationship as an overall structural fit, fares better not only in capturing the Bay model, but more generally in making sense of modelling practice.

References

Army Corps of Engineers. 1963. *Technical Report on Barriers: A Part of the Comprehensive Survey of San Francisco Bay and Tributaries, California*. Appendix H, Volume 1: Hydraulic Model Studies. San Francisco: Army Corps of Engineers.

Batterman, Robert. 2002. "Asymptotics and the Role of Minimal Models." *British Journal for the Philosophy of Science* 53 (1): 21-38.

Giere, Ronald N. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Giere, Ronald N. 1999a. *Science without Laws*. Chicago: University of Chicago Press.

Giere, Ronald N. 1999b. "Using Models to Represent Reality." In *Model-Based Reasoning in Scientific Discovery*, ed. Lorenzo Magnani, Nancy J. Nersessian, and Paul Thagard, 41-57. Springer Science & Business Media.

Giere, Ronald N. 2004. "How Models Are Used to Represent Reality." *Philosophy of Science* 71 (5): 742-752.

Giere, Ronald N. 2010. "An Agent-Based Conception of Models and Scientific Representation." *Synthese* 172 (2): 269–281.

Godfrey-Smith, Peter. 2006. "The Strategy of Model-based Science." *Biology and Philosophy* 21 (5): 725-740.

Haas, Leonard W. 1977. "The Effect of the Spring-Neap Tidal Cycle on the Vertical Salinity Structure of the James, York and Rappahannock Rivers, Virginia, U.S.A." *Estuarine and Coastal Marine Science* 5:485-496.

Huggins, Eugene. M., and Edward A. Schultz. 1967. “San Francisco Bay in A Warehouse.” *Journal of the IEST* 10 (5): 9-16.

Huggins, Eugene M., and Edward A. Schultz. 1973. “The San Francisco Bay and the Delta Model.” *California Engineer* 51 (3): 11-23.

Jackson, W. Turrentine, and Alan M. Peterson. 1977. *The Sacramento-San Joaquin Delta: The Evolution and Implementation of Water Policy*. Davis: California Water Resource Center, University of California.

Parker, Wendy. 2015. “Getting (even more) serious about similarity.” *Biology and Philosophy* 30 (2): 267-276.

Savenije, Hubert H. G. 2005. “*Salinity and Tides in Alluvial Estuaries*.” Elsevier Science.

Suárez, Mauricio. 2003. “Scientific Representation: against Similarity and Isomorphism.” *International Studies in the Philosophy of Science* 17 (3): 225-244.

Weisberg, Michael. 2012. “Getting Serious about Similarity.” *Philosophy of Science* 79 (5): 785-794.

Weisberg, Michael. 2013. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

PROBABILISTIC ACTUAL CAUSATION

LUKE FENTON-GLYNN

DEPARTMENT OF PHILOSOPHY, UNIVERSITY COLLEGE LONDON

GOWER STREET, LONDON, WC1E 6BT, U.K.

ABSTRACT. Actual (token) causes – e.g. Suzy's being exposed to asbestos – often bring about their effects – e.g. Suzy's suffering mesothelioma – probabilistically. I use probabilistic causal models to tackle one of the thornier difficulties for traditional accounts of probabilistic actual causation: namely probabilistic preemption.

Luke Fenton-Glynn

1. INTRODUCTION

Actual (token) causation is the relation that obtains when, for example, Suzy's being exposed to asbestos causes her to suffer mesothelioma. A number of theorists (e.g. Halpern and Pearl 2001, 2005; Hitchcock 2001, 2007; Weslake 2016) have deployed structural equations models (SEMs) in developing novel solutions to difficulties confronting traditional accounts of this relation. These theorists have focused on *deterministic* actual causation (DAC).[1] I draw on probabilistic causal models (PCMs) – analogues of deterministic SEMs – to provide an account of probabilistic actual causation (PAC). I don't attempt to show that my account can handle the full battery of test cases discussed in the literature. I simply demonstrate that it yields an elegant treatment of one very central case – probabilistic preemption – with a view to motivating further investigation of formal approaches to PAC.

2. PROBABILITY-RAISING

Probability-raising is central to the account developed here – as on traditional accounts of PAC.[2] To explain how I will understand that notion a bit of stage-setting is required.

I take the relata of the actual causal relation to be variable values. Adopting Goldszmidt and Pearl's (1992, 669–70) notation, $P(W = w|do(V = v))$ represents the probability for $W = w$ that *would* obtain if V were set to $V = v$ by an 'intervention' (Woodward 2005, 98). This is liable to diverge from the conditional probability $P(W = w|V = v)$: witness the difference between the probability of a storm *conditional* upon the barometer needle pointing toward the

[1] Cf. Halpern and Pearl (2005, 852); Hitchcock (2007, 498).

[2] Reichenbach (1971, 204); Suppes (1970); Lewis (1986, 175–84); Menzies (1989). The deficiencies of these accounts have been demonstrated by e.g. Salmon (1984, 192–202); Menzies (1996, 85–96); Hitchcock (2004).

word 'storm' and the probability of a storm if I had intervened upon the barometer needle to point it toward 'storm'.

Variable X taking value $X = x$ (rather than $X = x'$) raises the probability of $Y = y$ in the relevant sense iff:[3]

$$P(Y = y|do(X = x)) > P(Y = y|do(X = x')) \tag{1}$$

Appealing to interventionist probabilities means avoiding probability-raising relations between independent effects of a common cause, such as the barometer reading and the storm (cf. Lewis 1986, 178).

Probabilistic preemption cases illustrate that straightforward probability-raising is neither necessary nor sufficient for causation (Menzies 1989, 1996).

3. Probabilistic Preemption

The following example is inspired by Anscombe (1971).[4]

[3]Here and throughout, the probabilities (chances) should be taken to be those obtaining immediately after the interventions bringing about the variable values specified in the scope of the $do(\cdot)$ function have occurred (cf. Lewis 1986, 177).

[4]The probabilities involved (except the decision probabilities) are quantum and therefore objective and able underwrite causal relations. (If you're worried that the decision probabilities are not objective, the example could be complicated so that the decisions are made on the basis of outcomes of quantum measurements.) I find it plausible that the probabilities of many high level sciences are also objective (cf. e.g. Loewer 2001; Ismael 2009).

Luke Fenton-Glynn

(ProbPre) *Someone (neither you nor I) has connected a Geiger counter to a bomb so that the bomb will explode if the Geiger registers above a threshold reading. I place a place a chunk of U-232 (half-life = 68.9 years; decays by α-emission) near the Geiger. By chance, enough U-232 atoms decay within a short enough interval for the Geiger to reach the threshold reading so that the bomb explodes. Unbeknownst to me, you've been standing nearby observing. You have a chunk of Th-228 (half-life = 1.9 years; decays by α-emission), which contains many more atoms than my chunk of U-232. You've decided that you'll place your Th-228 near the Geiger iff I fail to place my U-232 near the Geiger. There's a negligible chance that you won't follow the course of action you've decided on. Seeing that I place my U-232 near the Geiger, you don't place your Th-228 near the Geiger.*[5]

Let M, D, Y, T, and E be binary variables which, respectively, take value 1 if the following things occur (and 0 otherwise): I place my U-232 near the Geiger; you decide to place your Th-228 near the Geiger iff I don't place my U-232 near the Geiger; you place your Th-228 near the Geiger; the threshold reading is reached; the bomb explodes.

My act ($M = 1$) was an actual cause of the explosion ($E = 1$). Yet plausibly the following inequality holds:

$$P(E = 1|do(M = 1)) < P(E = 1|do(M = 0)) \tag{2}$$

[5]The range of α-particles is 3-5 cm. Suppose that, for each of us, a decision to place our chunk 'near' the Geiger counter is a decision to place it $<$ 5cm away and a decision not to place it nearby is a decision to place it nowhere near ($\gg$ 5cm away).

That is, my placing my U-232 near the Geiger *lowers* the probability of the bomb exploding because it strongly lowers the probability of your placing your more potent Th-228 near the Geiger. Probability-raising is therefore unnecessary for actual causation.

Your decision ($D = 1$) was *not* an actual cause of the explosion, since you don't place your Th-228 near the Geiger. Yet provided there's some chance that $M = 0$, the following inequality holds:

$$P(E = 1|do(D = 1)) > P(E = 1|do(D = 0)) \tag{3}$$

Inequality (3) holds because your decision raises the probability that the bomb will still explode in the scenario in which $M = 0$.[6] Probability-raising is therefore insufficient for actual causation.

Actual causation therefore can't be identified with probability-raising. In developing a more nuanced analysis, it is helpful to appeal to PCMs.

4. PCMs

A PCM, $\mathscr{M}$, is a 5-tuple $\langle \mathscr{V}, \mathscr{C}, \Omega, \mathscr{F}, do(\cdot) \rangle$. $\mathscr{V}$ is a set of variables. Suppose $\mathscr{R}$ denotes a function from elements of $\mathscr{V}$ to sets of values: for all $V \in \mathscr{V}$, $\mathscr{R}(V)$ is the *range* of V. In Halpern and Pearl's (2005, 851–2) terminology, a formula $V_i = v_i$, for $V_i \in \mathscr{V}$ and $v_i \in \mathscr{R}(V)$, is a *primitive event*. $\mathscr{C}$ is the set of all those possible conjunctions of primitive

[6] $D = 0$ is multiply realizable: there is more than one alternative to the decision that you in fact make. E.g. you could decide that you will place your Th-228 near the Geiger no matter what, or that you will not do so no matter what. We can stipulate that the latter alternative is much more probable.

events, $V_1 = v_1 \& \ldots \& V_n = v_n$, such that $V_i \in \mathscr{V}$ and $v_i \in \mathscr{R}(V_i)$ and such that, for no pair of conjuncts $V_i = v_i$, $V_j = v_j$ is $V_i \equiv V_j$, and where no two elements of $\mathscr{C}$ differ *only* in the permutation of their conjuncts. Such a conjunction is denoted $\mathbf{V} = \mathbf{v}$ (primitive events and the null event are limiting cases of such conjunctions). Abusing notation, the fact that $v_i \in \mathscr{R}(V_i)$ for each primitive event $V_i = v_i$ in the conjunction $\mathbf{V} = \mathbf{v}$, is abbreviated $\mathbf{v} \in \mathscr{R}(\mathbf{V})$ and the set of variables that appear in $\mathbf{V} = \mathbf{v}$ is denoted $\mathbf{V}$.

Call a conjunction $\mathbf{V} = \mathbf{v}$ *maximal* if it contains a conjunct of the form $V_i = v_i$ for each $V_i \in \mathscr{V}$. Ω is the set of all maximal conjunctions of primitive events. $\mathscr{F}$ is a sigma algebra on Ω. Finally, $do(\cdot)$ is a function from elements of $\mathscr{C}$ to probability distributions on $\mathscr{F}$ (cf. Pearl 2009, 70, 110): for each element $\mathbf{V} = \mathbf{v}$ of $\mathscr{C}$, $P(\cdot|do(\mathbf{V} = \mathbf{v}))$ is the probability (chance) distribution on $\mathscr{F}$ that *would* obtain if interventions were performed to bring about $\mathbf{V} = \mathbf{v}$.

A PCM can be represented graphically by taking the variables in $\mathscr{V}$ as nodes and drawing a directed edge from V_i to V_j ($V_i, V_j \in \mathscr{V}$) iff, where $\mathbf{S} = \mathscr{V} \backslash V_i, V_j$, there is some assignment of values $\mathbf{s}' \in \mathscr{R}(\mathbf{S})$, some pair of values $v_i, v_i' \in \mathscr{R}(V_i)$ ($v_i \neq v_i'$) and some value $v_j \in \mathscr{R}(V_j)$ such that $P(V_j = v_j|do(V_i = v_i \& \mathbf{S} = \mathbf{s}')) \neq P(V_j = v_j|do(V_i = v_i' \& \mathbf{S} = \mathbf{s}'))$.

In constructing a PCM, $\mathscr{M}_{Pre}$, of **(ProbPre)** we might take the variable set to be $\mathscr{V}_{Pre} = \{D, M, Y, T, E\}$. The range of each variable in $\mathscr{V}_{Pre}$ is the pair $\{0, 1\}$. $\mathscr{C}_{Pre}$, Ω_{Pre}, and $\mathscr{F}_{Pre}$ are generated by $\mathscr{V}_{Pre}$ and $\mathscr{R}_{Pre}$ in the way described above. For each element of $\mathscr{C}_{Pre}$, the function $do(\cdot)$ returns the chance distribution on $\mathscr{F}_{Pre}$ that would obtain if interventions were performed to bring about that element of $\mathscr{C}_{Pre}$. The graph for $\mathscr{M}_{Pre}$ is given as figure 1.

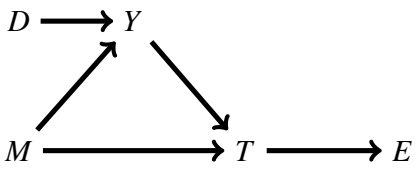


FIGURE 1

Probabilistic Actual Causation

A directed path in a graph is an ordered sequence of nodes, $\langle V_1, V_2, \ldots, V_n \rangle$, such that there is a directed edge from V_1 to V_2, and a directed edge from V_2 to $\ldots V_n$. $\langle M, Y, T, E \rangle$ is an example of a directed path in the graph of $\mathscr{M}_{Pre}$.

5. Appropriate Models

In Section 6, I provide a definition of what it is for $X = x$ (rather than $X = x'$) to count as an actual cause of $Y = y$ *relative to a PCM*. I then define a non-model-relativized notion of actual causation by saying that $X = x$ (rather than $X = x'$) counts as an actual cause of $Y = y$ *simpliciter* provided that $X = x$ (rather than $X = x'$) counts as an actual cause $Y = y$ relative to at least one *appropriate* PCM.[7] A similar strategy is commonly adopted by those analyzing DAC in terms of SEMs (Hitchcock 2001, 287, 2007, 503; Weslake 2016). This requires an account of 'appropriate' models.

Many of the criteria for an appropriate SEM for evaluating DAC carry over to PCMs, including the following three:

(Partition) For all $V \in \mathscr{V}$, the elements of $\mathscr{R}(V)$ should form a partition (Halpern and Hitchcock 2010, 397–8; Blanchard and Schaffer 2016)

(Independence) For no two variables $V, W \in \mathscr{V}$ should there be elements $v \in \mathscr{R}(V)$ and $w \in \mathscr{R}(W)$ such that the states of affairs represented by $V = v$ and $W = w$ are logically or metaphysically related (Hitchcock 2001, 287; Halpern and Hitchcock 2010, 397)

[7] As the parentheses indicate I define a *contrastive* relation of actual causation. Where variables are binary – as in $\mathscr{M}_{Pre}$ – this is inconsequential and I will typically suppress such parentheses. But it becomes important in cases of multi-valued variables (see Halpern and Pearl 2005, 859).

(Naturalness) For all $V \in \mathscr{V}$, $\mathscr{R}(V)$ should include only values that represent reasonably natural and intrinsic states of affairs. (Blanchard and Schaffer 2016)

The analysis of actual causation proposed below takes all and only values of distinct variables to be potential causal relata. (Partition) insures that we don't thereby miss actual causal relations because they obtain between the values of a single variable. (Independence) insures that we don't mistake stronger-than-causal relations for causal relations. (Naturalness) insures that unnatural or non-intrinsic states of affairs do not get counted as causes and effects (see Lewis 1986, 190, 263; Paul 2000, 245).[8]

A further condition is that a model is appropriate for evaluating whether $X = x$ is an actual cause of $Y = y$ in world θ only if it satisfies (Veridicality):

(Veridicality) For any conjunction $\mathbf{V} = \mathbf{v} \in \mathscr{C}$ taken as an input, the probability distribution $P(\cdot|do(\mathbf{V} = \mathbf{v}))$ yielded as an output by $do(\cdot)$ should be the *objective chance* distribution over $\mathscr{F}$ that would$_\theta$ result from interventions setting $\mathbf{V} = \mathbf{v}$. ('Would$_\theta$' indicates that what is required is that this counterfactual be true in θ.)

(Veridicality) is an analogue – for PCMs – of the requirement that SEMs encode only true counterfactuals (Hitchcock 2001, 287, 2007, 503).

In the DAC/SEMs literature another condition on model appropriateness is typically added:

(Serious Possibilities) $\mathscr{V}$ should not be such as to generate elements of Ω that represent possibilities "that we consider to be too remote" (Hitchcock 2001, 287;

[8] If *absences* are unnatural states of affairs (cf. Lewis 1986, 189–93), we might instead require that each variable have *at most one value* representing such a state of affairs.

cf. Woodward 2005, 86–91, Weslake 2016, Blanchard and Schaffer 2016).

We likely need this requirement too. A discussion of whether the vagueness and subjectivity thereby introduced is problematic would take us too far afield.[9] Still, it doesn't put the present account in any *worse* shape than its deterministic analogues. Moreover, traditional accounts of actual causation – which don't appeal to causal models – also stand in need of appeal to 'serious possibilities' (Woodward 2005, 86–8).

A final requirement – similar to one imposed in the DAC/SEM literature – for a model $\mathscr{M}$ to be an appropriate one for evaluating whether $X = x$ is an actual cause of $Y = y$ in world θ is:

(Stability) There is no model $\mathscr{M}^*$ (satisfying Partition, Independence, Naturalness, Veridicality, and Serious Possibilities) with a variable set $\mathscr{V}^*$ such that $\mathscr{V}^* \supset \mathscr{V}$ relative to which $X = x$ (rather than $X = x'$) is *not* an actual cause of $Y = y$. (Halpern and Hitchcock 2010, 394–5; Blanchard and Schaffer 2016; Halpern 2014; Hitchcock 2007, 503).

The idea is that an appropriate model is a sufficiently rich representation of causal reality that moving to a richer representation would not reveal an apparent actual causal relation to be spurious.[10]

The converse requirement – that a negative verdict about actual causation should not be overturned in a richer model – isn't needed. This is because actual causation (simpliciter) is defined in terms of actual causation relative to *at least one* appropriate model. A model relative verdict that $X = x$ is not an actual cause of $Y = y$ thus automatically fails to translate

[9] See Woodward (2005, 86–91).

[10] (Stability) renders the notion of an appropriate model relative to the causal claim being evaluated.

into a verdict that $X = x$ is not an actual cause (simpliciter) of $Y = y$ if there is a richer (and otherwise appropriate) model relative to which $X = x$ is an actual cause of $Y = y$.

We can now state a definition of actual causation in terms of appropriate PCMs that handles **(ProbPre)**.

6. PAC

Actual causation *simpliciter* is defined in terms of actual causation relative to an appropriate PCM. Model-relative actual causation is then defined.[11]

AC(S)
Where $x, x' \in \mathscr{R}(X)$ *and* $y \in \mathscr{R}(Y)$, $X = x$ *(rather than* $X = x'$*) is an* actual cause (simpliciter) *of* $Y = y$ *in world* θ *iff* $X = x$ *(rather than* $X = x'$*) is an actual cause of* $Y = y$ *relative to* at least one *model* $\mathscr{M}$ *(with* $X, Y \in \mathscr{V}$*) that is appropriate for evaluating whether* $X = x$ *(rather than* $X = x'$*) is an* actual cause (simpliciter) *of* $Y = y$ *in* θ.

[11] Those familiar with Halpern and Pearl's (2001, 2005) analyses of DAC are invited to see an analogy with **AC(M-R)**. **AC(M-R)** was partly inspired by thinking about how a counterpart of Halpern and Pearl's analysis might be developed that is adequate to the probabilistic case. Ultimately, I'm optimistic that an adequate account of DAC will fall out of an adequate account of PAC as the special case where all probabilities are 1 or 0. This is why my definitions take the definiendum to be 'actual cause' rather than 'probabilistic actual cause'.

Probabilistic Actual Causation

AC(M-R)

Where $x, x' \in \mathscr{R}(X)$ and $y \in \mathscr{R}(Y)$, $X = x$ (rather than $X = x'$) is an *actual cause* of Y = y relative to a model $\mathscr{M}$ (with $X, Y \in \mathscr{V}$) in world θ iff there is a partition $(\mathbf{Z}, \mathbf{W})$ of $\mathscr{V} \backslash X, Y$ and some setting $\mathbf{W} = \mathbf{w}'$ of the variables in $\mathbf{W}$ such that the $do(\cdot)$ function associated with $\mathscr{M}$ entails that, for all subsets $\mathbf{Z}'$ of $\mathbf{Z}$ (where, for each such subset, $\mathbf{Z}' = \mathbf{z}^*$ are the values that the variables in $\mathbf{Z}'$ have in θ):

(**IN**) $P(Y = y | do(X = x \& \mathbf{W} = \mathbf{w}' \& \mathbf{Z}' = \mathbf{z}^*)) > P(Y = y | do(X = x' \& \mathbf{W} = \mathbf{w}'))$

AC(M-R) counts $M = 1$ as an actual cause of $E = 1$ relative to $\mathscr{M}_{Pre}$ (and the world described in (**ProbPre**)). Consider the partition of $\mathscr{V}_{Pre} \backslash M, E$ such that $\mathbf{W} = \{D, Y\}$ and $\mathbf{Z} = \{T\}$. And consider the assignment $\{D = 1, Y = 0\}$ of values to the variables in $\mathbf{W}$. **AC(M-R)** is satisfied because (**IN**) holds for both subsets of $\mathbf{Z}$ ($\emptyset$ and $\{T\}$), as shown by (4) and (5):

(4) $P(E = 1 | do(M = 1 \& D = 1 \& Y = 0)) > P(E = 1 | do(M = 0 \& D = 1 \& Y = 0))$

(5) $P(E = 1 | do(M = 1 \& T = 1 \& D = 1 \& Y = 0)) > P(E = 1 | do(M = 0 \& D = 1 \& Y = 0))$

Inequality (4) indicates that my action raises the probability of the explosion *under the contingency* – i.e. *holding fixed* – that (you make your decision but) don't place your Th-228 near the Geiger. The existence of this *contingent* probability-raising reflects the fact that there is a path – $\langle M, T, E \rangle$ – along which $M = 1$ promotes $E = 1$ (because $M = 1$ raises the probability of $E = 1$ when we hold fixed the values of all variables off that path). It is the existence of

such a path – representing the process via which $M = 1$ produces $E = 1$ – that appears to drive our intuitions about actual causation in this case (cf. Hitchcock 2001).

Inequality (5) indicates that, again holding fixed $D = 1$ and $Y = 0$, the probability of $E = 1$ is higher if I place my U-232 near the Geiger *and the threshold reading is reached* than if I'd simply never placed my U-232 near the Geiger in the first place. As will be seen, this requirement ensures that, not only is there a potential process via which $M = 1$ threatens to bring about $E = 1$, but that process is complete.

Since **AC(M-R)** implies that $M = 1$ is an actual cause of $E = 1$ relative to $\mathscr{M}_{Pre}$, **AC(S)** yields the (correct) result that $M = 1$ is an actual cause (simpliciter) of $E = 1$ provided that $\mathscr{M}_{Pre}$ is appropriate. $\mathscr{M}_{Pre}$ *is* appropriate. Clearly it satisfies (Partition) and (Independence). It satisfies (Naturalness) because all of the states that its variables represent are reasonably natural. It was stipulated that the $do(\cdot)$ function associated with $\mathscr{M}_{Pre}$ is such that (Veridicality) is satisfied. $\mathscr{M}_{Pre}$ does not represent the sort of 'non-serious' possibility that (Serious Possibilities) is introduced to rule out (cf. Hitchcock 2001; Woodward 2005, 86–91).

Finally, (Stability) is satisfied because the causal process from my action to the explosion is complete. Holding fixed $Y = 0$, the probability of the explosion if $M = 1$ *and* part(s) of this process occur(s) is higher than the probability of the explosion if simply $M = 0$. Any variable (whose values represent reasonably natural states, form a partition, and are logically and metaphysically independent from the variables in $\mathscr{V}_{Pre}$) that might be added to $\mathscr{V}_{Pre}$ either represents part of this process or it doesn't. If it does, its actual value represents *the occurrence* of part of the process. So, if it is added to $\mathscr{V}_{Pre}$, including it in $\mathbf{Z}$ will not prevent (**IN**) from holding for all subsets $\mathbf{Z}'$ of $\mathbf{Z}$. If it doesn't, then adding it to $\mathscr{V}_{Pre}$, including it in $\mathbf{W}$, and holding it fixed at its actual value as part of the assignment $\mathbf{W} = \mathbf{w}'$ will not make a difference to the fact that (**IN**) holds for all subsets $\mathbf{Z}'$ of $\mathbf{Z}$, since holding fixed $Y = 0$ as part of $\mathbf{W} = \mathbf{w}'$ is already sufficient to ensure this.

Probabilistic Actual Causation

AC(M-R) gives the verdict that $D = 1$ is *not* an actual cause of $E = 1$ relative to $\mathscr{M}_{Pre}$. Consider the partition of $\mathscr{V}_{Pre}\backslash D, E$ such that $\mathbf{W} = \{M\}$ and $\mathbf{Z} = \{Y, T\}$. Observe that:

$$P(E = 1|do(D = 1\&M = 0)) > P(E = 1|do(D = 0\&M = 0)) \tag{6}$$

And:

$$P(E = 1|do(D = 1\&M = 1)) > P(E = 1|do(D = 0\&M = 1)) \tag{7}$$

Thus, whichever possible value we hold fixed M at, the probability of $E = 1$ is higher if $D = 1$ than if $D = 0$. So $D = 1$ contingently raises the probability of $E = 1$.[12] That's because there's a path – $\langle D, Y, E\rangle$ – along which $D = 1$ promotes $E = 1$.

AC(M-R) nevertheless entails that $D = 1$ is *not* an actual cause of $E = 1$ relative to $\mathscr{M}_{Pre}$. Consider the subset $\{Y\}$ of $\mathbf{Z}$, and observe that:

$$P(E = 1|do(D = 1\&Y = 0\&M = 0)) \leq P(E = 1|do(D = 0\&M = 0)) \tag{8}$$

And:

$$P(E = 1|do(D = 1\&Y = 0\&M = 1)) \leq P(E = 1|do(D = 0\&M = 1)) \tag{9}$$

That is, whichever possible value we hold fixed M at, the probability of the explosion is no higher if you make your decision *but don't place your Th-228 near the Geiger* than if you'd

[12] The obtaining of just one of (6) or (7) would suffice to show this.

never made that decision in the first place. Thus (**IN**) does not hold for every subset of **Z** for this partition of variables no matter what values we assign to the variables in **W**. This reflects the fact that, because you didn't place your Th-228 near the Geiger, there is no complete causal process by which your decision produces the explosion. Your non-placement of your Th-228 'neutralizes' the danger of your decision causing the explosion.

Is there an alternative partition (**W**, **Z**) of $\mathscr{V}_{Pre}$ and assignment $\mathbf{W} = \mathbf{w}'$ such that (**IN**) holds for all subsets $\mathbf{Z}'$ of **Z**? (There need only be *one* for **AC(M-R)** to be satisfied.) There isn't. Assigning Y to **W** instead of **Z** won't help, since the value of Y 'screens off' D from E. So, where $Y \in \mathbf{W}$, no assignment $\mathbf{W} = \mathbf{w}'$ will be such that, holding fixed $\mathbf{W} = \mathbf{w}'$, the probability of $E = 1$ is higher when $D = 1$ (and the variables in $\emptyset \subseteq \mathbf{Z}$ are set to their actual values) than when $D = 0$. So (**IN**) doesn't hold for all subsets $\mathbf{Z}'$ of **Z** for any such partition.

On the other hand, if we leave Y in **Z** and also assign M to **Z**, then there are no variables in **W** to hold fixed. Now consider the subset $\{Y\}$ of **Z**, and observe that:[13]

$$P(E = 1|do(D = 1 \& Y = 0)) \leq P(E = 1|do(D = 0)) \tag{10}$$

So, with M assigned to **Z** it remains the case that (**IN**) doesn't hold for all subsets of **Z**.

So there's no partition of $\mathscr{V}_{Pre} \backslash D, E$ such that (**IN**) is satisfied for all subsets of **Z** when we consider $D = 1$ as a putative cause of $E = 1$. **AC(M-R)** therefore doesn't count $D = 1$ as an actual cause of $E = 1$ relative to $\mathscr{M}_{Pre}$.

But for **AC(S)** to count $D = 1$ as an actual cause of $E = 1$ *simpliciter*, there need only be one appropriate model relative to which **AC(M-R)** counts $D = 1$ as an actual cause of $E = 1$. Is there such a model? There isn't. Suppose a candidate such model includes Y. Because D is only relevant to E because of its relevance to Y, the value of Y 'screens off' the value of D

[13] Note: the fact that $Y = 0$ *due to an intervention* doesn't make $M = 1$ more likely.

from that of E. This means that, if Y is included in **W** in the partition (**W**, **Z**) of the model's variable set and held fixed (either at 1 or 0) as part of the assignment $\mathbf{W} = \mathbf{w}'$, then (**IN**) won't be satisfied for the empty subset of **Z**. Alternatively, if Y is included in **Z** then, no matter what other variables are included in the model and assigned to **W**, (**IN**) won't be satisfied for the subset $\{Y\}$ of **Z**. Specifically, because $D = 1$ only threatens to bring about $E = 1$ because it threatens to bring about $Y = 1$, no matter what we hold fixed by inclusion on both sides of (**IN**), the probability of $E = 1$ is no higher if $D = 1$ *and* $Y = 0$ than if simply $D = 0$.

So **AC(M-R)** doesn't count $D = 1$ as an actual cause of $E = 1$ relative to any appropriate model with Y in its variable set. This means that any otherwise appropriate model relative to which $D = 1$ *is* an actual cause of $E = 1$ can be expanded to a model in which $D = 1$ *isn't* an actual cause of $E = 1$ simply by the addition of Y. Provided the expanded model is appropriate, the original model violates (Stability) and is inappropriate. So **AC(S)** will correctly not count $D = 1$ as an actual cause *simpliciter* of $E = 1$.

Since the values of Y form a partition and represent natural states of affairs, (Partition) and (Naturalness) will be satisfied by the expanded model if they were satisfied by the original model. With regard to (Veridicality), it should be noted that there are multiple ways of expanding the original model via the addition of Y, each associated with a different $do(\cdot)$ function from elements of $\mathscr{C}^*$ to probability distributions over $\mathscr{F}^*$ (where $\mathscr{C}^*$ and $\mathscr{F}^*$ are generated by the expanded variable set in the way described in Section 4). In looking for an apt expanded model, we just select the one with the $do(\cdot)$ function that returns the objective chances on $\mathscr{F}^*$ that *would* obtain as a result of interventions bringing about the various elements of $\mathscr{C}^*$. With regard to (Serious Possibilities) note that, given your decision, your placing *and* your not placing your Th-228 near the Geiger are both salient possibilities in

(ProbPre). So it doesn't seem that the expanded model could represent any non-serious possibilities if the original model doesn't. (Independence) is a little trickier. Might not the original model include a variable whose values are logically or metaphysically related to those of Y? Given that the variables in the original model are assumed to satisfy (Partition) it seems that any variable logically or metaphysically related to Y – e.g. Y', which takes value $Y' = 0$ if you don't place your Th-228 near the Geiger, $Y' = 1$ if you place it 2.5-5cm from the Geiger, and $Y' = 2$ if you place it 0-2.5cm from the Geiger – will also be such that its actual value neutralizes the threat of $D = 1$ bringing about $E = 1$, so that **AC(M-R)** is not satisfied in the original model. The exception to this would be if the original model included a variable that represents a gerrymandered states of affairs – e.g. Y'', which takes value $Y'' = 1$ if you place your Th-228 near the Geiger *or* Obama is US president, and $Y'' = 0$ otherwise – in which case the original model will violate (Naturalness).

7. Conclusion

Drawing upon PCMs, an account of PAC has been given that gives a correct treatment of probabilistic preemption on intuitive grounds. Traditional accounts of PAC misdiagnose this central test case (Menzies, 1989, 1996; Hitchcock 2004). Examination of whether PCMs can help tackle some of the other outstanding problems of PAC is warranted.

REFERENCES

Anscombe, E. (1971). *Causality and Determination*. Cambridge: CUP.

Blanchard, T. and J. Schaffer (2016). Cause without Default. In H. Beebee, C. Hitchcock, and H. Price (Eds.), *Making a Difference*. Oxford: OUP.

Goldszmidt, M. and J. Pearl (1992). Rank-Based Systems. In *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, San Mateo, CA, pp. 661–672. Morgan Kaufmann.

Halpern, J. Y. (2014). Appropriate Causal Models and Stability of Causation. In *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning*, Palo Alto, CA, pp. 198–207. AAAI Press.

Halpern, J. Y. and C. Hitchcock (2010). Actual Causation and the Art of Modeling. In R. Dechter, H. Geffner, and J. Y. Halpern (Eds.), *Heuristics, Probability and Causality*, pp. 383–406. London: College Publications.

Halpern, J. Y. and J. Pearl (2001). Causes and Explanations: A Structural-Model Approach. Part I: Causes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, pp. 194–202. Morgan Kaufmann.

Halpern, J. Y. and J. Pearl (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal for the Philosophy of Science 56*, 843–87.

Hitchcock, C. (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy 98*, 194–202.

Hitchcock, C. (2004). Do All and Only Causes Raise the Probabilities of Effects? In J. Collins, N. Hall, and L. Paul (Eds.), *Causation and Counterfactuals*, pp. 403–417. Cambridge, MA: MIT Press.

Hitchcock, C. (2007). Prevention, Preemption, and the Principle of Sufficient Reason. *Philosophical Review 116*, 495–532.

Ismael, J. (2009). Probability in Deterministic Physics. *Journal of Philosophy 106*, 89–108.

Lewis, D. (1986). *Philosophical Papers*, Volume 2. Oxford: Oxford University Press.

Loewer, B. (2001). Determinism and Chance. *Studies in History and Philosophy of Science Part B 32*, 609–620.

Menzies, P. (1989). Probabilistic Causation and Causal Processes: A Critique of Lewis. *Philosophy of Science 56*, 642–663.

Menzies, P. (1996). Probabilistic Causation and the Pre-emption Problem. *Mind 105*, 85–117.

Paul, L. (2000). Aspect Causation. *Journal of Philosophy 97*, 235–256.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (Second ed.). Cambridge: CUP.

Reichenbach, H. (1971). *The Direction of Time*. Mineola, NY: Dover.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Suppes, P. (1970). *A Probabilistic Theory of Causality, Acta Philosophica Fennica*. Amsterdam: North-Holland.

Weslake, B. (2016). A Partial Theory of Actual Causation. Forthcoming in *British Journal for the Philosophy of Science*.

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford: OUP.

When Journal Editors Play Favorites*

Remco Heesen†

June 28, 2016

Abstract

Should editors of scientific journals practice triple-blind reviewing? I consider two arguments in favor of this claim. The first says that insofar as editors' decisions are affected by information they would not have had under triple-blind review, an injustice is committed against certain authors. I show that even well-meaning editors would commit this wrong and I endorse this argument.

The second argument says that insofar as editors' decisions are affected by information they would not have had under triple-blind review, it will negatively affect the quality of published papers. I distinguish between two kinds of biases that an editor might have. I show that one of them has a positive effect on quality and the other a negative one, and that the combined effect could be either positive or negative. Thus I do not endorse the second argument in general. However, I do endorse this argument for certain fields, for which I argue that the positive effect does not apply.

*Thanks to Kevin Zollman and Liam Bright for valuable comments and discussion. This work was partially supported by the National Science Foundation under grant SES 1254291.

†Department of Philosophy, Baker Hall 161, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA. Email: rheesen@cmu.edu

1 Introduction

Journal editors occupy an important position in the scientific landscape. By making the final decision on which papers get published in their journal and which papers do not, they have a significant influence on what work is given attention and what work is ignored in their field (Crane 1967).

In this paper I investigate the following question: should the editor be informed about the identity of the author when she is deciding whether to publish a particular paper? Under a single- or double-blind reviewing procedure, the editor has access to information about the author, whereas under a triple-blind reviewing procedure she does not. So in other words the question is: should journals practice triple-blind reviewing?

Two kinds of arguments have been given in favor of triple-blind reviewing. One focuses on the treatment of the author by the editor. On this kind of argument, revealing identity information to the editor will lead the editor to (partially) base her judgment on irrelevant information (such as the gender of the author, or whether or not the editor is friends with the author). This harms the author, and is thus bad.

The second kind of argument focuses on the effect on the journal and its readers. Again, the idea is that the editor will base her judgment on identity information if given the chance to do so. But now the further claim is that as a result the journal will accept worse papers. After all, if a decision to accept or reject a paper is influenced by the editor's biases, this suggests that a departure has been made from a putative "objectively correct" decision. This harms the readers of the journal, and is thus bad.

Here I provide a philosophical discussion of the reviewing procedure to assess these arguments. I distinguish between two different ways the editor's judgment may be affected if the author's identity is revealed to her. First, the editor may treat authors she knows differently from authors she does not know. Second, the editor may treat authors differently based on their membership of some group (e.g., gender bias). My discussion focuses on the

following three claims.

My first claim is that the first kind of differential treatment the editor may display (based on whether she knows a particular author) actually benefits rather than harms the readers of the journal. This benefit is the result of a reduction in editorial uncertainty about the quality of submitted papers when she knows their authors. I construct a model to show in a formally precise way how such a benefit might arise—surprisingly, no assumption that the scientists the editor knows are somehow "better scientists" is required—and I cite empirical evidence that such a benefit indeed does arise. However, this benefit only applies in certain fields. I argue that in other fields (in particular, mathematics and the humanities) no significant reduction of uncertainty—and hence no benefit to the readers—occurs (section 2).

My second claim is that either kind of differential treatment the editor may display (based on whether she knows authors or based on bias against certain groups) harms authors. I argue that any instance of such differential treatment constitutes an epistemic injustice in the sense of Fricker (2007) against the disadvantaged author. If the editor is to be (epistemically) just, she should prevent such differential treatment, which can be done through triple-blind reviewing. So I endorse an argument of the first of the two kinds I identified above: triple-blind reviewing is preferable because not doing so harms authors (section 3).

My third claim is that whether differential treatment also harms the journal and its readers depends on a number of factors. Differential treatment by the editor based on whether she knows a particular author may benefit readers, whereas differential treatment based on bias against certain groups may harm them. Whether there is an overall benefit or harm depends on the strength of the editor's bias, the relative sizes of the different groups, and other factors, as I illustrate using the model. As a result I do not in general endorse the second kind of argument, that triple-blind reviewing is preferable because readers of the journal are harmed otherwise. However, I do endorse

this argument for fields like mathematics and the humanities, where I claim that the benefits of differential treatment (based on uncertainty reduction) do not apply (section 4).

Note that, in considering the ethical and epistemic effects of triple-blind reviewing, a distinction is made between the effects on the author and the effects on the readers of the journal. This reflects a growing understanding that in order to study the social epistemology of science, what is good for an individual inquirer must be distinguished from what is good for the wider scientific community (Kitcher 1993, Strevens 2003, Mayo-Wilson et al. 2011).

Zollman (2009) has studied the effects of different editorial policies on the number of papers published and the selection criteria for publication, but he does not focus specifically on the editor's decisions and the uncertainty she faces. Economists have studied models in which editor decisions play an important role (Ellison 2002, Faria 2005, Besancenot et al. 2012), but they have not distinguished between papers written by scientists the editor knows and papers by scientists unknown to her, and neither have they been concerned with biases the editor may be subject to. And some other economists have done empirical work investigating the differences between papers with and without an author-editor connection (Laband and Piette 1994, Medoff 2003, Smith and Dombrowski 1998, more on this later), but they do not provide a model that can explain these differences. This paper thus fills a gap in the literature.

2 A Model of Editor Uncertainty

As I said in the introduction, journal editors have a certain measure of power in a scientific community because they decide which papers get published.[1] An editor could use this discretionary power to the benefit of her friends or

[1]Different journals may have different policies, such as one in which associate editors make the final decision for papers in their (sub)field. Here, I simply define "the editor" to be whomever makes the final decision whether to publish a particular paper.

colleagues, or to promote certain subfields or methodologies over others. This phenomenon has been called *editorial favoritism*. If anecdotal evidence is to be believed, this phenomenon is widespread. Some systematic evidence of favoritism exists as well. Bailey et al. (2008a,b) find that academics believe editorial favoritism to be fairly prevalent, with a nonnegligible percentage claiming to have perceived it firsthand. Laband (1985) and Piette and Ross (1992) find that, controlling for citation impact and various other factors, papers whose author has a connection to the journal editor are allocated more journal pages than papers by authors without such a connection.[2]

In this paper, I refer to the phenomenon that editors are more likely to accept papers from authors they know than papers from authors they do not know as *connection bias*.

Academics tend to disapprove of this behavior (Sherrell et al. 1989, Bailey et al. 2008a,b). In both of the studies by Bailey et al., in which subjects were asked to rate the seriousness of various potentially problematic behaviors by editors and reviewers, this disapproval was shown (using a factor analysis) to be part of a general and strong disapproval of "selfish or cliquish acts" in the peer review process. Thus it appears that the reason for the disapproval of editors publishing papers by their friends and colleagues is that it shows the editor acting on private interests, rather than displaying the disinterestedness that is the norm in science (Merton 1942).

On the other hand, if connection bias was a serious worry for authors, one would expect this to be a major consideration for them in choosing where to submit their papers (i.e., submit to journals where they know the editor), but Ziobrowski and Gibler (2000) find that this is not the case.[3]

[2]Here, page allocation is used as a proxy for journal editors' willingness to push the paper. The more obvious variable to use here would be whether or not the paper is accepted for publication. Unfortunately, there are no empirical studies which measure the influence of a relationship between the author and the editor on acceptance decisions directly. Presumably this is because information about rejected papers is usually not available in these kinds of studies.

[3]In particular, authors who know an editor and thus could expect to profit from con-

Moreover, despite working scientists' disapproval, there is some evidence that connection bias improves the overall quality of accepted papers (Laband and Piette 1994, Medoff 2003, Smith and Dombrowski 1998). Does that mean scientists are misguided in their disapproval?

As indicated in the introduction, I distinguish between the effects of editors' biases on the authors of scientific papers on the one hand, and the effects on the readers of scientific journals on the other hand. In this section, I use a formal model to show that these two can come apart: connection bias may negatively affect scientists as authors while positively affecting scientists as readers. Note that in this section I focus only on connection bias. Subsequent sections consider other biases.

Consider a simplified scientific community consisting of a set of scientists. Each scientist produces a paper and submits it to the community's only journal which has one editor.

Some papers are more suitable for publication than others. I assume that this suitability for publication can be measured on a single numerical scale. For convenience I call this the *quality* of the paper. However, I remain neutral on how this notion should be interpreted, e.g., as an objective measure of the epistemic value of the paper (which is perhaps an aggregate of multiple relevant criteria), or as the number of times the paper would be cited in future papers if it was published, or as the average subjective value each member of the scientific community would assign to it if they read it.[4]

nection bias would find knowing the editor and the composition of the editorial board more generally to be important factors in deciding where to submit, contrary to Ziobrowski and Gibler's evidence (these factors are ranked twelfth and sixteenth in importance in a list of sixteen factors that might influence the decision where to submit). Similarly, authors who do not know an editor would find a lack of (perceived) connection bias and the composition of the editorial board to be important factors, but these rank only seventh and twelfth in importance in Ziobrowski and Gibler's study. In a similar survey by Mackie (1998, chapter 4), twenty percent of authors indicated that knowing the editor and/or her preferences is an important consideration in deciding where to submit a paper.

[4]For more on potential difficulties with interpreting the notion of quality, see Bright (2015).

Crucially, the editor does not know the quality of the paper at the time it is submitted. The aim of this section is to show how uncertainty about quality can lead to connection bias. To make this point as starkly as possible, I assume that the editor cares only about quality, i.e., she makes an estimate of the quality of a paper and publishes those and only those papers whose quality estimate is high.

Let q_i be the quality of the paper submitted by scientist i. Since there is uncertainty about the quality, q_i is modeled as a random variable. Since some scientists are more likely to produce high quality papers than others, the mean μ_i of this random variable may be different for each scientist. I assume that quality follows a normal distribution with fixed variance: $q_i \mid \mu_i \sim N(\mu_i, \sigma^2_{qu})$.

The assumptions of normality and fixed variance are made primarily to keep the mathematics simple. Below I make similar assumptions on the distribution of average quality in the scientific community and the distribution of reviewers' estimates of the quality of a paper. I see no reason to expect the results I present below to be different when any of these assumptions are changed.

If the editor knows scientist i, she has some prior information on the average quality of scientist i's work. This is reflected in the model by assuming that the editor knows the value of μ_i. For scientists she does not know, the editor is uncertain about the average quality of their work. All she knows is the distribution of average quality in the larger scientific community, which I also assume to be normal: $\mu_i \sim N(\mu, \sigma^2_{sc})$.

Note that I assume the scientific community to be homogeneous: the scientific community is split in two groups (those known by the editor and those not known by the editor) but average paper quality follows the same distribution in both groups. If I assumed instead that scientists known by the editor write better papers on average the results would be qualitatively similar to those I present below. If scientists known by the editor write worse

papers on average this would affect my results. However, since most journal editors are relatively central figures in their field (Crane 1967), this would be an implausible assumption except perhaps in isolated cases.

The editor's prior beliefs about the quality of a paper submitted by some scientist i reflects this difference in information. If she knows the scientist she knows the value of μ_i, and so her prior is $\pi(q_i \mid \mu_i) \sim N(\mu_i, \sigma^2_{qu})$. If the editor does not know scientist i she only knows the distribution of μ_i, rather than its exact value. Integrating out the uncertainty over μ_i yields a prior $\pi(q_i) \sim N(\mu, \sigma^2_{qu} + \sigma^2_{sc})$ for the quality of scientist i's paper.

When the editor receives a paper she sends it out for review. In the context of this model, the main purpose of the reviewer's report is to provide an estimate of the quality of the paper. But, I assume, even after reading the paper its quality cannot be established with certainty. Thus the reviewer's estimate r_i of the quality q_i is again a random variable. I assume that the reviewer's report is unbiased, i.e., its mean is the actual quality q_i of the paper. Once again I use a normal distribution to reflect the uncertainty: $r_i \mid q_i \sim N(q_i, \sigma^2_{rv})$.[5]

The editor uses the information from the reviewer's report to update her beliefs about the quality of scientist i's paper. I assume that she does this by Bayes conditioning. Thus, her posterior beliefs about the quality of the paper are $\pi(q_i \mid r_i)$ if she does not know the author, and $\pi(q_i \mid r_i, \mu_i)$ if she does.

The posterior distributions are themselves normal distributions whose

[5]The reviewer's report could reflect the opinion of a single reviewer, or the averaged opinion of multiple reviewers. The editor could even act as a reviewer herself, in which case the report reflects her findings which she has to incorporate in her overall beliefs about the quality of the paper. The assumption I make in the text can be used to cover any of these scenarios, as long as a given journal is fairly consistent in the number of reviewers used. If the number of reviewers is frequently different for different papers (and in particular when this difference correlates with the existence or absence of a connection between editor and author) the assumption of a fixed variance in the reviewer's report is unrealistic because a report from multiple reviewers may be thought to give more accurate information (reducing the variance) than a report from a single reviewer.

mean is a weighted average of r_i and the prior mean, as given in proposition 1 (for a proof, see DeGroot 2004, section 9.5, or any other textbook that covers Bayesian statistics).

Proposition 1.

$$\pi(q_i \mid r_i) \sim N\left(\mu_i^U, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right),$$
$$\pi(q_i \mid r_i, \mu_i) \sim N\left(\mu_i^K, \frac{\sigma_{qu}^2\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right),$$

where

$$\mu_i^U = \frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\mu,$$
$$\mu_i^K = \frac{\sigma_{qu}^2}{\sigma_{qu}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\mu_i.$$

When does the editor choose to publish a paper? Here I assume that she publishes any paper whose posterior mean is above some threshold q^*. So a paper written by a scientist unknown to the editor is published if $\mu_i^U > q^*$ and a paper written by a scientist known to the editor is published if $\mu_i^K > q^*$. This corresponds to being at least 50% confident that the paper's quality is above the threshold. Other standards could be used (risk-averse standards might require more than 50% confidence that the paper is above some threshold, while risk-loving standards might require less; in these cases the threshold value needs to be adapted to keep the total number of accepted papers constant) but for my purposes here it does not much matter.

Now compare the probability that the paper of an arbitrary scientist i unknown to the editor is published to the probability that the paper of an arbitrary scientist known by the editor is published. For this purpose it is useful to determine the probability distribution of the posterior means (see appendix A for proofs of this and subsequent results).

Proposition 2. *The posterior means are normally distributed, with* $\mu_i^U \sim N(\mu, \sigma_U^2)$ *and* $\mu_i^K \sim N(\mu, \sigma_K^2)$. *Here,*

$$\sigma_U^2 = \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} \qquad \textit{and} \qquad \sigma_K^2 = \frac{\sigma_{qu}^4 + \sigma_{sc}^2(\sigma_{qu}^2 + \sigma_{rv}^2)}{\sigma_{qu}^2 + \sigma_{rv}^2}.$$

Moreover, if $\sigma_{sc}^2 > 0$ *and* $\sigma_{rv}^2 > 0$, *then* $\sigma_U^2 < \sigma_K^2$.

The main result of this section, which establishes the existence of connection bias in the model, is a consequence of proposition 2. It says that the editor is more likely to publish a paper written by an arbitrary author she knows than a paper written by an arbitrary author she does not know, whenever $q^* > \mu$ (for any positive value of σ_{sc}^2 and σ_{rv}^2). Since $q^* = \mu$ would mean that exactly half of all papers gets published, the condition amounts to a requirement that the journal's acceptance rate is less than 50%. This is true of most reputable journals in most fields (physics being a notable exception). When acceptance rates are above 50% editorial favoritism is also much less of a concern in the first place.

Theorem 3. *If* $q^* > \mu$, $\sigma_{sc}^2 > 0$, *and* $\sigma_{rv}^2 > 0$, *the acceptance probability for authors known to the editor is higher than the acceptance probability for authors unknown to the editor, i.e.,* $\Pr\left(\mu_i^K > q^*\right) > \Pr\left(\mu_i^U > q^*\right)$.

Theorem 3 shows that in the model I presented, any journal with an acceptance rate lower than 50% will be seen to display connection bias. Thus I have established the surprising result that an editor who cares only about the quality of the papers she publishes may end up publishing more papers by her friends and colleagues than by scientists unknown to her, even if her friends and colleagues are not, as a group, better scientists than average.

Why does this surprising result hold? The theorem follows immediately from proposition 2, which says that the distribution of μ_i^U is less "spread out" than the distribution of μ_i^K ($\sigma_U^2 < \sigma_K^2$). This happens because μ_i^U is a

weighted average of μ and r_i, keeping it relatively close to the overall mean μ compared to μ_i^K, which is a weighted average of μ_i and r_i (which tend to differ from μ in the same direction).

Because the editor treats papers by authors she knows differently from papers by authors she does not know, authors unknown to the editor are arguably harmed. I pick up this point in section 3 and argue that this constitutes an epistemic injustice against those authors.

What I have shown so far is that an editor who uses information about the average quality of papers produced by scientists she knows in her acceptance decisions will find that scientists she knows produce on average more papers that meet her quality threshold. This is a subjective statement: the editor believes that more papers by scientists she knows meet her threshold. Does this translate into an objective effect? That is, does the extra information the editor has available about scientists she knows allow her to publish better papers from them than from scientists she does not know?

In order to answer this question I need to compare the average quality of accepted papers. More formally, I want to compare the expected value of the quality of a paper, conditional on meeting the publication threshold, given that the author is either known to the editor or not.

Proposition 4. *If $\sigma_{sc}^2 > 0$, and $\sigma_{rv}^2 > 0$, the average quality of accepted papers from authors known to the editor is higher than the average quality of accepted papers from authors unknown to the editor, i.e., $\mathbb{E}[q_i \mid \mu_i^K > q^*] > \mathbb{E}[q_i \mid \mu_i^U > q^*]$.*

Proposition 4 shows that the editor can use the extra information she has about scientists she knows to improve the average quality of the papers published in her journal. In other words, the surprising result is that the editor's connection bias actually benefits rather than harms the readers of the journal. It is thus fair to say that, in the model, the editor can use her connections to "identify and capture high-quality papers", as Laband and

Piette (1994) suggest.[6]

To what extent does this show that the connection bias observed in reality is the result of editors capturing high-quality papers, as opposed to editors using their position of power to help their friends? At this point the model is seen to yield an empirical prediction. If connection bias is (primarily) due to capturing high-quality papers, the quality of papers by authors the editor knows should be higher than average, as shown in the model. If, on the other hand, connection bias is (primarily) a result of the editor accepting for publication papers written by authors she knows even though they do not meet the quality standards of the journal, then the quality of papers by authors the editor knows should (presumably) be lower than average.

If subsequent citations are a good indication of the quality of a paper,[7] a simple regression can test whether accepted papers written by authors with an author-editor connection have a higher or a lower average quality than papers without such a connection. This empirical test has been carried out a number of times, and the results univocally favor the hypothesis that editors use their connections to improve the quality of published papers (Laband and Piette 1994, Smith and Dombrowski 1998, Medoff 2003).

Note that in the above results, nothing depends on the sizes of the variances σ^2_{qu}, σ^2_{sc}, and σ^2_{rv}. This is because these results are qualitative. The variances do matter when the acceptance rate and average quality of papers are compared quantitatively. For example, reducing σ^2_{rv} (making the reviewer's report more accurate) makes the differences in the acceptance rate and average quality of papers smaller.

[6]This result applies to connection bias only. Below I consider other biases the editor might have, which yields more nuanced conclusions.

[7]Recall that I have remained neutral on how the notion of quality should be interpreted. If quality is simply defined as "the number of citations this paper would get if it were published" the connection between quality and citations is obvious. Even on other interpretations of quality, citations have frequently been viewed as a good proxy measure (Cole and Cole 1967, 1968, Medoff 2003). This practice has been defended by Cole and Cole (1971) and Clark (1957, chapter 3), and criticized by Lindsey (1989) and Heesen (forthcoming).

Note also that the results depend on the assumption that σ^2_{sc} and σ^2_{rv} are positive. What is the significance of these assumptions?

If $\sigma^2_{rv} = 0$, i.e., if there is no variance in the reviewer's report, the reviewer's report describes the quality of the paper with perfect accuracy. In this case the "extra information" the editor has about authors she knows is not needed, and so there is no difference in acceptance rate or average quality based on whether the editor knows the author. But it seems unrealistic to expect reviewer's reports to be this accurate.

If $\sigma^2_{sc} = 0$ there is either no difference in the average quality of papers produced by different authors, or learning the identity of the author does not tell the editor anything about the expected quality of that scientist's work. In this case there is no value to the editor (with regard to determining the quality of the submitted paper) in learning the identity of the author. So here also there is no difference in acceptance rate or average quality based on whether the editor knows the author.

Under what circumstances should the identity of the author be expected to tell the editor something useful about the quality of a submitted paper? This seems to be most obviously the case in the lab sciences. The identity of the author, and hence the lab at which the experiments were performed, can increase or decrease the editor's confidence that the experiments were performed correctly, including all the little checks and details that are impossible to describe in such a paper. In a scientific paper, "[a]s long as the conclusions depend at least in part on the results of some experiment, the reader must rely on the author's (and perhaps referee's) testimony that the author really performed the experiment exactly as claimed, and that it worked out as reported" (Easwaran 2009, p. 359).

But in other fields, in particular mathematics and some or all of the humanities, there is no need to rely on the author's reputation. This is because in these fields the paper itself is the contribution, so it is possible to judge papers in isolation of how or by whom they were created. Easwaran

(2009) discusses this in detail for mathematics, and briefly (in his section 4) for philosophy. And in fact there exists a norm that this is how they should be judged: "Papers will rely only on premises that the competent reader can be assumed to antecedently believe, and only make inferences that the competent reader would be expected to accept on her own consideration." (Easwaran 2009, p. 354).

Arguably then, the advantage (see theorem 3 and proposition 4) conferred by revealing identity information about the author to the editor applies only in certain fields. The relevant fields are those where part of the information in the paper is conferred on the authority of testimony, in particular those where experimental results are reported. Even in those fields, of course, what is being testified is supposed to be reproducible by the reader. But this is still different from the case in mathematics and the humanities, where a careful reading of a paper itself constitutes a reproduction of its argument. In these latter fields there is no relevant information to be learned from the identity of the author (i.e., $\sigma^2_{sc} = 0$), or, at least, the publishing norms in these fields suggest that their members believe this to be the case.

3 Bias As an Epistemic Injustice

The previous section discussed a formal model of editorial uncertainty about paper quality. The first main result, theorem 3, established the existence of connection bias in this model: authors known by the editor are more likely to see their paper accepted than authors unknown to the editor. The second main result, proposition 4, showed that connection bias benefits the readers of the journal by improving the average quality of accepted papers.

Despite the benefit to the readers, I claim that authors are harmed by connection bias. In this section I argue that an instance of connection bias constitutes an *epistemic injustice* in the sense of Fricker (2007). Then I argue that the editor is likely to display other biases as well, and that instances of

these also constitute epistemic injustices.

The type of epistemic justice that is relevant here is *testimonial injustice*. Fricker (2007, pp. 17–23) defines a testimonial injustice as a case where a speaker suffers a credibility deficit for which the hearer is ethically and epistemically culpable, rather than being due to innocent error.

Testimonial injustices may arise in various ways. Fricker is particularly interested in what she calls "the central case of testimonial injustice" (Fricker 2007, p. 28). This kind of injustice results from a *negative identity-prejudicial stereotype*, which is defined as follows:

> A widely held disparaging association between a social group and one or more attributes, where this association embodies a generalization that displays some (typically, epistemically culpable) resistance to counter-evidence owing to an ethically bad affective investment. (Fricker 2007, p. 35)

Because the stereotype is widely held, it produces *systematic* testimonial injustice: the relevant social group will suffer a credibility deficit in many different social spheres.

Applying this to the phenomenon of connection bias, it is clear that this is not an instance of the central case of testimonial injustice. This would entail that there is some negative stereotype associated with scientists unknown to the editor, as a group, which is not normally the case. So I set the central case aside (I return to it below) and focus on the question whether connection bias can produce (non-central cases of) testimonial injustice.

Suppose scientist i and scientist i' tend to produce papers of the same quality, which is above average in the population ($\mu_i = \mu_{i'} > \mu$). Suppose further that the actual papers they have produced on this occasion are of the same quality ($q_i = q_{i'}$) and have received similar reviewer reports ($r_i = r_{i'}$). If scientist i is not known to the editor, but scientist i' is, then the paper

written by scientist i' is likely to be evaluated more highly by the editor.[8] If the publication threshold q^* is somewhere in between the two evaluations then only scientist i' will have her paper accepted.

In this example, the scientists produced papers of equal quality that were evaluated differently. So scientist i suffers a credibility deficit. This deficit is not due to innocent error, as it would be if, e.g., random variation led to different reviewer reports (i.e., $r_i < r_{i'}$). The deficit is also not due to the editor's use of generally reliable information about the two scientists, as it would be if there was a genuine difference in the average quality of the papers they produce (i.e., $\mu_i < \mu_{i'}$).

Is this credibility deficit suffered by scientist i ethically and epistemically culpable on the part of the editor? On the one hand, as I stressed in section 2, the editor is simply making maximal use of the information available to her. It just so happens that she has more information about scientists she knows than about others. But that is hardly the editor's fault: she cannot be expected to know everyone's work. Is it incumbent upon her to get to know the work of every scientist who submits a paper?

This may well be too much to ask. But an alternative option is to remove all information about the authors of submitted papers. This can be done by using a triple-blind reviewing procedure, in which the editor does not know the identity of the author, and hence is prevented from using information about scientists she knows in her evaluation. Using such a procedure, at least all scientists are treated equally: any scientist who writes a paper of a given quality has the same chance of seeing that paper accepted.

So a credibility deficit occurs which harms scientist i: her paper is rejected. Moreover, it harms her specifically as an epistemic agent: the rejection of the paper reflects a judgment of the quality of her scientific work. And

[8]The editor's posterior mean for the quality of scientist i's paper is μ_i^U and her posterior mean for scientist i''s paper is $\mu_{i'}^K = \mu_i^K$, with $\mu_i^U < \mu_{i'}^K$ whenever $\sigma_{sc}^2(r_i - \mu_i) < (\sigma_{qu}^2 + \sigma_{rv}^2)(\mu_i - \mu)$. The claim in the text is then justified by the fact that $\Pr(\sigma_{sc}^2(r_i - \mu_i) < (\sigma_{qu}^2 + \sigma_{rv}^2)(\mu_i - \mu) \mid \mu_i > \mu) > 1/2$, assuming $\sigma_{sc}^2 > 0$ and $\sigma_{rv}^2 > 0$.

this harm could have been prevented by the editor by using a triple-blind reviewing procedure.

I conclude that the editor is ethically and epistemically culpable for this credibility deficit, and hence a testimonial injustice is committed against scientist i. However, one may insist that it cannot be the case that the editor is committing a wrong simply in virtue of using relevant information that is available to her. An evidentialist in particular may say that it cannot possibly be an epistemic wrong to take into account all relevant information.

I disagree, for the reasons just given, but I need not insist on this point. Even if it is granted that the editor does not commit an injustice by using the information that is available to her, the end result is still that scientist i is harmed as an epistemic agent. She has produced a paper of equal quality to scientist i''s, and yet it is not published.

Moreover, the presence of scientist i' is irrelevant. Any time a paper from an author unknown to the editor is rejected which would have been accepted had the editor known the author (all else being equal), that author is harmed. So even if one insists that differential editorial treatment resulting from connection bias is not culpable on the part of the editor, connection bias still harms authors whenever it influences acceptance decisions.

In the model of section 2, and the above discussion, I assumed that connection bias is the only bias journal editors display. The literature on implicit bias suggests that this is not true. For example, "[i]f submissions are not anonymous to the editor, then the evidence suggests that women's work will probably be judged more negatively than men's work of the same quality" (Saul 2013, p. 45). Evidence for this claim is given by Wennerås and Wold (1997), Valian (1999, chapter 11), Steinpreis et al. (1999), Budden et al. (2008), and Moss-Racusin et al. (2012).[9] So women scientists are at

[9]These citations show that the work of women in academia is undervalued in various ways. None of them focus specifically on editor evaluations, but they support Saul's claim unless it is assumed that journal editors as a group are significantly less biased than other academics.

a disadvantage simply because of their gender identity. Similar biases exist based on other irrelevant aspects of scientists' identity, such as race or sexual orientation (see Lee et al. 2013, for a critical survey of various biases in the peer review system). As Crandall (1982, p. 208) puts it: "The editorial process has tended to be run as an informal, old-boy network which has excluded minorities, women, younger researchers, and those from lower-prestige institutions".

I use *identity bias* to refer to these kinds of biases. Any time a paper is rejected because of identity bias (i.e., the paper would have been accepted if the relevant part of the author's identity had been different, all else being equal), a testimonial injustice occurs for the same reasons outlined above. Moreover, here the editor is culpable for having these biases.

Unlike instances resulting from connection bias, testimonial injustices resulting from identity bias can be instances of the central case of testimonial injustice, in which the credibility deficit results from a negative identity-prejudicial stereotype. The evidence suggests that negative identity-prejudicial stereotypes affect the way people (not just men) judge women's work, even when the person judging does not consciously believe in these stereotypes. Moreover, those who think highly of their ability to judge work objectively and/or are primed with objectivity are affected more rather than less (Uhlmann and Cohen 2007, Stewart and Payne 2008, p. 1333). Similar claims plausibly hold for biases based on race or sexual orientation. Biases based on academic affiliation are not usually due to negative identity-prejudicial stereotypes, as these do not generally affect other aspects of the scientist's life.

So both connection bias and identity bias are responsible for injustices against authors. This is one way to spell out the claim that authors are harmed when journal editors do not use a triple-blind reviewing procedure. This constitutes the first kind of argument for triple-blind reviewing which I mentioned in the introduction, and which I endorse based on these consid-

erations.

4 The Effect of Bias on Quality

The second kind of argument I mentioned in the introduction claims that failing to use triple-blind reviewing harms the journal and its readers, because it would lower the average quality of accepted papers. In section 2 I argued that connection bias actually has the opposite effect: it increases average quality. In this section I complicate the model to include identity bias.

Recall that the editor displays identity bias if she is more or less likely to publish papers from a certain group of scientists based on some aspect of their identity, e.g., their gender. I incorporate this in the model by assuming the editor consistently undervalues members of one group (and overvalues the others). More precisely, she believes the average quality of papers produced by any scientist i from the group she is biased against to be lower than it really is by some constant quantity ε. Conversely, the average quality of papers written by any scientist not belonging to this group is raised by δ.[10] So the editor has a different prior for the two groups; I use π_A to denote her prior for the quality of papers written by scientists she is biased against, and π_F for her prior for scientists she is biased in favor of.

As before, the editor may be familiar with a given scientist's work (i.e., she knows the average quality of that scientist's papers) or not. So there are now four groups. If scientist i is known to the editor and belongs to the stigmatized group the editor's prior distribution on the quality of scientist i's paper is $\pi_A(q_i \mid \mu_i) \sim N(\mu_i - \varepsilon, \sigma^2_{qu})$. If scientist i is known to the editor but is not in the stigmatized group the prior is $\pi_F(q_i \mid \mu_i) \sim N(\mu_i + \delta, \sigma^2_{qu})$. If

[10]This is a simplifying assumption: one could imagine having biases against multiple groups of different strengths, or biases whose strength has some random variation, or biases which intersect in various ways (Collins and Chepp 2013, Bright et al. 2016). However, the assumption in the main text suffices to make the point I want to make. It should be fairly straightforward to extend my results to more complicated cases like the ones just described.

scientist i is not known to the editor and is in the stigmatized group the prior is $\pi_A(q_i) \sim N(\mu - \varepsilon, \sigma_{qu}^2 + \sigma_{sc}^2)$. And if scientist i is not known to the editor and not in the stigmatized group the prior is $\pi_F(q_i) \sim N(\mu + \delta, \sigma_{qu}^2 + \sigma_{sc}^2)$.[11]

The next few steps in the development are analogous to that in section 2. After the reviewer's report comes in the editor updates her beliefs about the quality of the paper, yielding the following posterior distributions.

Proposition 5.

$$\pi_A(q_i \mid r_i, \mu_i) \sim N\left(\mu_i^{KA}, \frac{\sigma_{qu}^2 \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right),$$
$$\pi_F(q_i \mid r_i, \mu_i) \sim N\left(\mu_i^{KF}, \frac{\sigma_{qu}^2 \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right),$$
$$\pi_A(q_i \mid r_i) \sim N\left(\mu_i^{UA}, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right),$$
$$\pi_F(q_i \mid r_i) \sim N\left(\mu_i^{UF}, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right),$$

where

$$\mu_i^{KA} = \mu_i^K - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \qquad \mu_i^{KF} = \mu_i^K + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2},$$
$$\mu_i^{UA} = \mu_i^U - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \qquad \mu_i^{UF} = \mu_i^U + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}.$$

As before, the paper is published if the posterior mean (μ_i^{KA}, μ_i^{KF}, μ_i^{UA}, or μ_i^{UF}) exceeds the threshold q^*. The respective distributions of the posterior

[11]Note that I assume that the editor displays bias against scientists in the stigmatized group regardless of whether she knows them or not. Under a reviewing procedure that is not triple-blind, the editor learns at least the name and affiliation of any scientist who submits a paper. This information is usually sufficient to determine with reasonable certainty the scientist's gender. So at least for gender bias it seems reasonable to expect the editor to display bias even against scientists she does not know. Conversely, because negative identity-prejudicial stereotypes can work unconsciously, it does not seem reasonable to expect that the editor can withhold her bias from scientists she knows.

means determine how likely this is. These distributions are given in the next proposition.

Proposition 6. *The posterior means are normally distributed, with*

$$\begin{aligned}
\mu_i^{KA} &\sim N\left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2\right), \\
\mu_i^{KF} &\sim N\left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2\right), \\
\mu_i^{UA} &\sim N\left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2\right), \\
\mu_i^{UF} &\sim N\left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2\right).
\end{aligned}$$

This yields the within-group acceptance rates and the unsurprising result that the editor is less likely to publish papers by scientists she is biased against.

Theorem 7. *If $\varepsilon > 0$, $\delta > 0$, and $\sigma_{rv}^2 > 0$, the acceptance probability for authors the editor is biased against is lower than the acceptance probability for authors the editor is biased in favor of (keeping fixed whether or not the editor knows the author). That is,*

$$\Pr\left(\mu_i^{KA} > q^*\right) < \Pr\left(\mu_i^{KF} > q^*\right) \text{ and } \Pr\left(\mu_i^{UA} > q^*\right) < \Pr\left(\mu_i^{UF} > q^*\right).$$

Theorem 7 establishes the existence of identity bias in the model: authors that are subject to a negative identity-prejudicial stereotype are less likely to see their paper accepted than authors who are not. As I argued in section 3, whenever a paper is rejected due to identity bias this constitutes a testimonial injustice against the author.

Now I turn my attention to the effect that identity bias has on the average quality of accepted papers. In the current version of the model there is both

connection bias and identity bias. Connection bias has been shown to have a positive effect on average quality (see section 2). Whether the net effect of connection bias and identity bias is positive or negative depends on various parameters, as I illustrate below.

The benchmark for judging the average quality of accepted papers under a procedure subject to connection bias and identity bias is a *triple-blind reviewing procedure* under which the editor is not informed of the identity of the scientist. As a result, she is both unable to use information about the average quality of a given scientist's papers and unable to display bias against scientists based on their identity.

Under this triple-blind procedure, the editor's prior distribution for the quality of any submitted paper is $\pi(q_i) \sim N(\mu, \sigma_{qu}^2 + \sigma_{sc}^2)$, i.e., the prior I used in section 2 when the author was unknown to the editor. Hence, under this procedure, the posterior is $\pi(q_i \mid r_i)$, the posterior mean is $\mu_i^U \sim N(\mu, \sigma_U^2)$, the probability of acceptance is $\Pr(\mu_i^U > q^*)$ and the average quality of accepted papers is $\mathbb{E}[q_i \mid \mu_i^U > q^*]$.

In contrast, I refer to the reviewing procedure that is subject to connection bias and identity bias as the *non-blind procedure*. The overall probability that a paper is accepted under the non-blind procedure depends on the relative sizes of the four groups. I use p_{KA} to denote the fraction of scientists known to the editor that she is biased against, p_{KF} for the fraction known to the editor that she is biased in favor of, p_{UA} for unknown scientists biased against, and p_{UF} for unknown scientists biased in favor of. These fractions are nonnegative and sum to one.

Let A_i denote the event that scientist i's paper is accepted under the non-blind procedure. The overall probability of acceptance under this procedure is

$$\Pr(A_i) = p_{KA} \Pr\left(\mu_i^{KA} > q^*\right) + p_{KF} \Pr\left(\mu_i^{KF} > q^*\right) + p_{UA} \Pr\left(\mu_i^{UA} > q^*\right) + p_{UF} \Pr\left(\mu_i^{UF} > q^*\right).$$

The average quality of accepted papers can then be written as $\mathbb{E}[q_i \mid A_i]$. I want to compare $\mathbb{E}[q_i \mid A_i]$ to $\mathbb{E}[q_i \mid \mu_i^U > q^*]$, the average quality of accepted papers under a triple-blind procedure.[12]

In the remainder of this section I assume that the editor's biases are such that she believes the average quality of all submitted papers to be equal to μ. In other words, her bias against the stigmatized group is canceled out on average by her bias in favor of those not in the stigmatized group, weighted by the relative sizes of those groups:

$$(p_{KA} + p_{UA})\,\varepsilon = (p_{KF} + p_{UF})\,\delta.$$

I use the above equation to fix the value of δ, reducing the number of free parameters by one. The equation amounts to a kind of commensurability requirement for the two procedures because it guarantees that the editor perceives the average quality of submitted papers to be the same regardless of whether or not a triple-blind procedure is used.

As far as I can tell there are no interesting general conditions on the parameter values that determine whether the non-blind procedure or the triple-blind procedure will lead to a higher average quality of accepted papers. The question I will explore now, using some numerical examples, is how biased the editor needs to be for the epistemic costs of her identity bias to outweigh the epistemic benefits resulting from connection bias.

In order to generate numerical data values have to be chosen for the

[12]Expressions for $\Pr(A_i)$ and $\mathbb{E}[q_i \mid A_i]$ using only the parameter values and standard functions are given in lemma 11 in appendix A. These expressions are used to generate the numerical results below.

parameters. First I set $\mu = 0$ and $q^* = 2$. Since quality is an interval scale in this model, these choices are arbitrary. For the variances σ^2_{qu}, σ^2_{sc}, and σ^2_{rv}, I choose a "small" and a "large" value (1 and 4 respectively).

For the sizes of the four groups, I assume that there is no correlation between whether the editor knows an author and whether the editor has a bias against that author (so, e.g., the percentage of women among scientists the editor knows is equal to the percentage of women among scientists the editor does not know). I consider two cases for the editor's identity bias: either she is biased against half the set of authors (and so biased in favor of the other half) or the group she is biased against is a 30 % minority.[13] Similarly, I consider the case in which the editor knows half of all scientists submitting papers, and the case in which the editor knows 30 % of them.

As a result, there are 32 possible settings of the parameters (2^3 choices for the variances times 2^2 choices for the group sizes). Whether the triple-blind procedure or the non-blind procedure is epistemically preferable depends on the value of ε (and the value of δ determined thereby).

It follows from proposition 4 that when $\varepsilon = 0$ the non-blind procedure helps rather than harms the readers of the journal by increasing average quality relative to the triple-blind procedure. If ε is positive but relatively small, this remains true, but when ε is relatively big, the non-blind procedure harms the readers. This is because the average quality of published papers under the non-blind procedure decreases continuously as ε increases (I do not prove this, but it is easily checked for the 32 cases I consider).

The interesting question, then, is where the turning point lies. How big does the editor's bias need to be in order for the negative effects of identity bias on quality to cancel out the positive effects of connection bias?

[13]Bruner and O'Connor (forthcoming) note that certain dynamics in academic life can lead to identity bias against groups as a result of the mere fact that they are a minority. Here I consider both the case where the stigmatized group is a minority (and is possibly stigmatized as a result of being a minority, as Bruner and O'Connor suggest) and the case where it is not (and so presumably the negative identity-prejudicial stereotype has some other source).

I determine the value of ε for which the average quality of published papers under the non-blind procedure and the triple-blind procedure is the same for each of the 32 cases. But reporting these numbers directly does not seem particularly useful, as ε is measured in "quality points" which do not have a clear interpretation outside of the model.

To give a more meaningful interpretation of these values of ε as measuring "size of bias", I calculate the average rate of acceptance of papers from authors the editor is biased against and the average rate of acceptance of papers from authors the editor is biased in favor of.[14] The difference between these numbers gives an indication of the size of the editor's bias: it measures (in percentage points, abbreviated pp) how many more papers the editor accepts from authors she is biased in favor of, compared to those she is biased against.

This difference is reported for the 32 cases in figure 1. To provide a sense of scale for these numbers, I plot them against the acceptance rate that the triple-blind procedure would have for those values of the parameters, i.e., $\Pr(\mu_i^U > q^*)$.

Already with this small sample of 32 cases, a large variation of results can be observed. I illustrate this by looking at two cases in detail.

First, suppose that $\sigma_{qu}^2 = \sigma_{sc}^2 = 1$ and $\sigma_{rv}^2 = 4$. In this extreme case the triple-blind procedure has an acceptance rate as low as 0.72 %. If the groups are all of equal size ($p_{KA} = p_{KF} = p_{UA} = p_{UF} = 1/4$) then under the non-blind procedure the acceptance rate for authors the editor is biased in favor of needs to be as much as 2.66 pp higher than the acceptance rate for authors the editor is biased against, in order for the average quality under

[14] These are calculated without regard for whether the editor knows the author or not. In particular, the rate of acceptance for authors the editor is biased against is

$$\frac{p_{KA}\Pr\left(\mu_i^{KA} > q^*\right) + p_{UA}\Pr\left(\mu_i^{UA} > q^*\right)}{p_{KA} + p_{UA}}, \text{ and } \frac{p_{KF}\Pr\left(\mu_i^{KF} > q^*\right) + p_{UF}\Pr\left(\mu_i^{UF} > q^*\right)}{p_{KF} + p_{UF}}$$

is the rate of acceptance for authors the editor is biased in favor of.

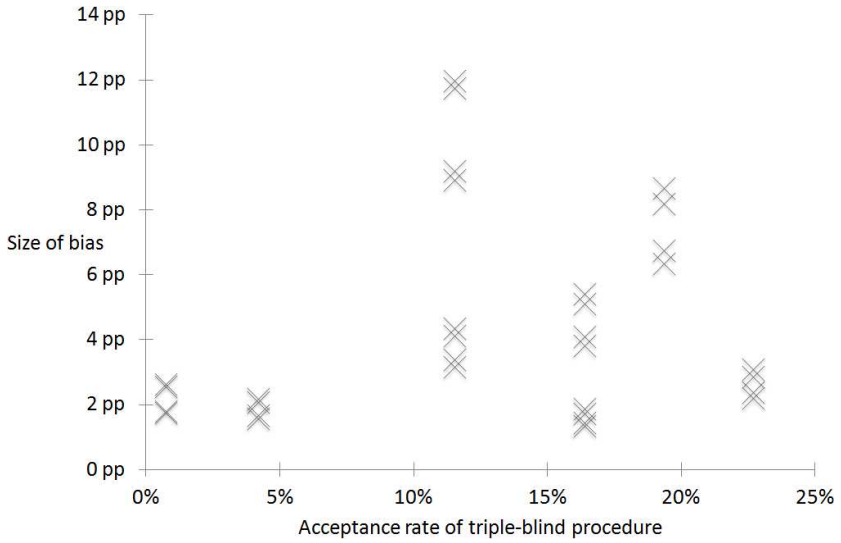


Figure 1: The minimum size of the editor's bias such that the quality costs of the non-blind procedure outweigh its benefits (given as a percentage point difference in acceptance rates), in 32 cases, plotted as a function of the acceptance rate of the corresponding triple-blind procedure.

the two procedures to be equal. Clearly a 2.66 pp bias is very large for a journal that only accepts less than 1 % of papers. If the bias is any less than that there is no harm to the readers in using the non-blind procedure.

Second, suppose that $\sigma_{qu}^2 = \sigma_{sc}^2 = 4$ and $\sigma_{rv}^2 = 1$. Then the triple-blind procedure has an acceptance rate of 22.66 %. If, moreover, the editor knows relatively few authors ($p_{KA} = p_{KF} = 0.15$, $p_{UA} = p_{UF} = 0.35$) then the acceptance rate for authors the editor is biased in favor of needs to be only 2.23 pp higher than the acceptance rate for authors the editor is biased against, in order for the quality costs of the non-blind procedure to outweigh its benefits. For a journal accepting about 23 % of papers that means that even if the identity bias of the editor is relatively mild the journal's readers are harmed if the non-blind procedure is used.

Based on these results, and the fact that the parameter values are unlikely to be known in practice, it is unclear whether the non-blind procedure

or the triple-blind procedure will lead to a higher average quality of published papers for any particular journal.[15] So in general it is not clear that an argument that the non-blind procedure harms the journal's readers can be made. At the same time, a general argument that the non-blind procedure helps the readers is not available either. Given this, I am inclined to recommend a triple-blind procedure for all journals because not doing so harms the authors.

If there was reason to believe that the editor's bias was very small, there might be a case for the non-blind procedure using considerations of average quality. Based on the empirical evidence I cited in section 3, it seems unlikely that any editor could make such a case convincingly today. But if identity bias were someday to be eliminated or severely mitigated, this question may be worth revisiting.

So far I have argued in this section that in the presence of the positive effect of connection bias on quality, the net effect of connection bias and identity bias on quality is unclear. But I argued in section 2 that the positive effect of connection bias may only exist in certain fields. In fields where papers rely partially on the author's testimony there is value in knowing the identity of the author. But in other fields such as mathematics and some of the humanities testimony is not taken to play a role—the paper itself constitutes the contribution to the field—and so arguably there is no value in knowing the identity of the author.

In those fields, then, there is no quality benefit from connection bias, but there is still a quality cost from identity bias. So here the strongest case for the triple-blind procedure emerges, as the non-blind procedure harms both authors and readers.

[15]Note that the evidence collected by Laband and Piette (1994) does not help settle this question, as they do not directly compare the triple-blind and the non-blind procedure. Their evidence supports a positive epistemic effect of connection bias, but not a verdict on the overall epistemic effect of triple-blinding.

5 Conclusion

In this paper I have considered two types of arguments for triple-blind review: one based on the consequences for the author and one based on the consequences for the readers of the journal.

I have argued that the non-blind procedure introduces differential treatment of scientific authors. In particular, editors are more likely to publish papers by authors they know (connection bias, theorem 3) and less likely to publish papers by authors they apply negative identity-prejudicial stereotypes to (identity bias, theorem 7). Whenever a paper is rejected as a result of one of these biases an epistemic injustice (in the sense of Fricker 2007) is committed against the author. This is an argument in favor of triple-blinding based on consequences for the author.

From the readers' perspective the story is more mixed. Generally speaking connection bias has a positive effect on the quality of published papers and identity bias a negative one. Thus whether the readers are better off under the triple-blind procedure depends on how exactly these effects trade off, which is highly context-dependent, or so I have argued. This yields a more nuanced view than that suggested by either Laband and Piette (1994), who focus only on connection bias, or by the argument for triple-blinding based on the consequences for the readers, which focuses only on identity bias.

However, in mathematics and some of the humanities there is arguably no positive quality effect from connection bias, as knowing about an author's other work is not taken to be relevant (Easwaran 2009). So here the negative effect of identity bias is the only relevant consideration from the readers' perspective. In this situation, considerations concerning the consequences for the author and considerations concerning the consequences for the readers point in the same direction: in favor of triple-blind review.

A The Acceptance Probability and the Average Quality of Papers

Proposition 2. *$\mu_i^U \sim N(\mu, \sigma_U^2)$ and $\mu_i^K \sim N(\mu, \sigma_K^2)$. Moreover, $\sigma_U^2 < \sigma_K^2$ whenever $\sigma_{sc}^2 > 0$ and $\sigma_{rv}^2 > 0$.*

Proof. First consider the distribution of r_i. Since $r_i \mid q_i \sim N(q_i, \sigma_{rv}^2)$, $q_i \mid \mu_i \sim N(\mu_i, \sigma_{qu}^2)$, and $\mu_i \sim N(\mu, \sigma_{sc}^2)$, it follows that $r_i \mid \mu_i \sim N(\mu_i, \sigma_{qu}^2 + \sigma_{rv}^2)$ and $r_i \sim N(\mu, \sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)$.

The latter can be used straightforwardly to determine the distribution of μ_i^U. Since $r_i - \mu \sim N(0, \sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)$ it follows that

$$\frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}(r_i - \mu) \sim N\left(0, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right) \sim N\left(0, \sigma_U^2\right).$$

The result follows because μ is a constant and

$$\mu_i^U = \frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\mu = \frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}(r_i - \mu) + \mu.$$

Determining the distribution of μ_i^K is slightly trickier because there are two random variables involved: r_i and μ_i. As noted above, $r_i \mid \mu_i \sim N(\mu_i, \sigma_{qu}^2 + \sigma_{rv}^2)$. Thus, writing $X_i = \frac{\sigma_{qu}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}(r_i - \mu_i)$,

$$X_i \mid \mu_i \sim N\left(0, \frac{\sigma_{qu}^4}{\sigma_{qu}^2 + \sigma_{rv}^2}\right).$$

Since

$$\mu_i^K = \frac{\sigma_{qu}^2}{\sigma_{qu}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\mu_i = X_i + \mu_i$$

it remains to determine the convolution of X_i and μ_i. This can be done using

the moment-generating function and the law of total expectation. Recall that the moment-generating function of an $N(m, s^2)$ distribution is given by $M(t) = \exp\{mt + \frac{1}{2}s^2t^2\}$. So the moment-generating function of μ_i^K is

$$\begin{aligned}
\mathbb{E}[\exp\{t\mu_i^K\}] &= \mathbb{E}[\exp\{t(X_i + \mu_i)\}] \\
&= \mathbb{E}[\mathbb{E}[\exp\{tX_i + t\mu_i\} \mid \mu_i]] \\
&= \mathbb{E}[\exp\{t\mu_i\}\mathbb{E}[\exp\{tX_i\} \mid \mu_i]] \\
&= \exp\left\{0t + \frac{1}{2}\frac{\sigma_{qu}^4}{\sigma_{qu}^2 + \sigma_{rv}^2}t^2\right\}\mathbb{E}[\exp\{t\mu_i\}] \\
&= \exp\left\{\frac{1}{2}\frac{\sigma_{qu}^4}{\sigma_{qu}^2 + \sigma_{rv}^2}t^2 + \mu t + \frac{1}{2}\sigma_{sc}^2t^2\right\} \\
&= \exp\left\{\mu t + \frac{1}{2}\frac{\sigma_{qu}^4 + \sigma_{sc}^2(\sigma_{qu}^2 + \sigma_{rv}^2)}{\sigma_{qu}^2 + \sigma_{rv}^2}t^2\right\},
\end{aligned}$$

which is exactly the moment-generating function of the desired normal distribution.

Finally, note that

$$\begin{aligned}
\sigma_U^2 &= \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2(\sigma_{qu}^2 + \sigma_{rv}^2)}{(\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)(\sigma_{qu}^2 + \sigma_{rv}^2)}, \\
\sigma_K^2 &= \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2(\sigma_{qu}^2 + \sigma_{rv}^2) + \sigma_{sc}^2\sigma_{rv}^4}{(\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)(\sigma_{qu}^2 + \sigma_{rv}^2)}.
\end{aligned}$$

So $\sigma_U^2 < \sigma_K^2$ whenever $\sigma_{sc}^2 > 0$ and $\sigma_{rv}^2 > 0$ (and $\sigma_U^2 = \sigma_K^2$ otherwise, assuming the expressions are well-defined in that case). □

Theorem 3. $\Pr\left(\mu_i^K > q^*\right) > \Pr\left(\mu_i^U > q^*\right)$ *if* $q^* > \mu$, $\sigma_{sc}^2 > 0$, *and* $\sigma_{rv}^2 > 0$.

Proof. It follows from proposition 2 that

$$\Pr\left(\mu_i^K > q^*\right) = 1 - \Phi\left(\frac{q^* - \mu}{\sigma_K}\right) \text{ and } \Pr\left(\mu_i^U > q^*\right) = 1 - \Phi\left(\frac{q^* - \mu}{\sigma_U}\right),$$

where Φ is the distribution function (or cumulative density function) of a standard normal distribution. Since Φ is (strictly) increasing in its argument, and $\sigma_K > \sigma_U$ by proposition 2, the theorem follows immediately. □

In order to prove proposition 4 a number of intermediate results are needed.

Lemma 8.

$$\mathbb{E}[q_i \mid \mu_i^U > q^*] = \mathbb{E}[\mu_i^U \mid \mu_i^U > q^*],$$
$$\mathbb{E}[q_i \mid \mu_i^K > q^*] = \mathbb{E}[\mu_i^K \mid \mu_i^K > q^*].$$

Proof. Because μ_i^U is simply an (invertible) transformation of r_i, it follows that

$$q_i \mid \mu_i^U \sim q_i \mid r_i \sim N\left(\mu_i^U, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right).$$

The distribution of $q_i \mid \mu_i^K$ is a little trickier to find, because μ_i^K is a linear combination of two random variables, r_i and μ_i, and it is not obvious that learning μ_i^K is as informative as learning both r_i and μ_i. But using the known distributions of $q_i \mid \mu_i$ and $\mu_i^K \mid q_i, \mu_i$ and integrating out μ_i it can be shown that

$$q_i \mid \mu_i^K \sim q_i \mid r_i, \mu_i \sim N\left(\mu_i^K, \frac{\sigma_{qu}^2\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right).$$

The important point here is that $\mathbb{E}[q_i \mid \mu_i^x] = \mu_i^x$ both for $x = U$ and $x = K$.

Now the law of total expectation can be used to establish that

$$\mathbb{E}[q_i \mid \mu_i^x > q^*] = \mathbb{E}[\mathbb{E}[q_i \mid \mu_i^x] \mid \mu_i^x > q^*] = \mathbb{E}[\mu_i^x \mid \mu_i^x > q^*],$$

for $x = U, K$. □

Let $X \sim N(\mu, \sigma^2)$ be a normally distributed random variable. Then $X \mid X > a$ follows a *left-truncated normal distribution*, with left-truncation point a. As a result of lemma 8 I am interested in the mean of left-truncated normal distributions. According to, e.g., Johnson et al. (1994, chapter 13, section 10.1), this mean can be expressed as

$$\mathbb{E}[X \mid X > a] = \mu + \sigma R\left(\frac{a-\mu}{\sigma}\right). \tag{1}$$

Here

$$R(x) = \frac{\phi(x)}{1 - \Phi(x)}$$

for all $x \in \mathbb{R}$, where ϕ is the probability density function of the standard normal distribution, and Φ is its distribution function. R is the inverse of what is known in the literature (e.g., Gordon 1941) as *Mills' ratio*.

It follows from the definitions that $R(x) > 0$ for all $x \in \mathbb{R}$ and that

$$R'(x) = R(x)^2 - xR(x). \tag{2}$$

Proposition 9 (Gordon (1941))**.** *For all $x > 0$, $R(x) < \frac{x^2+1}{x}$.*

Proposition 9 can be used to establish the next result.

Proposition 10. *If $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\mu, s^2)$ with $s > \sigma > 0$ then $\mathbb{E}[Y \mid Y > a] > \mathbb{E}[X \mid X > a]$.*

Proof. It suffices to show that the derivative $\frac{\partial}{\partial \sigma}\mathbb{E}[X \mid X > a]$ is positive for all $\sigma > 0$. Differentiating equation (1) (using equation (2)) yields

$$\frac{\partial}{\partial\sigma}\mathbb{E}[X \mid X > a] = \left(\left(\frac{a-\mu}{\sigma}\right)^2 + 1\right) R\left(\frac{a-\mu}{\sigma}\right) - \frac{a-\mu}{\sigma} R\left(\frac{a-\mu}{\sigma}\right)^2.$$

Since $R\left(\frac{a-\mu}{\sigma}\right) > 0$, $\frac{\partial}{\partial\sigma}\mathbb{E}[X \mid X > a] > 0$ if and only if

$$\left(\frac{a-\mu}{\sigma}\right)^2 + 1 - \frac{a-\mu}{\sigma} R\left(\frac{a-\mu}{\sigma}\right) > 0.$$

This is true whenever $\frac{a-\mu}{\sigma} \leq 0$ because then both terms in the sum are positive. Proposition 9 guarantees that it is true whenever $\frac{a-\mu}{\sigma} > 0$ as well. □

Proposition 4. $\mathbb{E}[q_i \mid \mu_i^K > q^*] > \mathbb{E}[q_i \mid \mu_i^U > q^*]$ *whenever* $\sigma_{sc}^2 > 0$, *and* $\sigma_{rv}^2 > 0$.

Proof. By lemma 8,

$$\mathbb{E}[q_i \mid \mu_i^U > q^*] = \mathbb{E}[\mu_i^U \mid \mu_i^U > q^*],$$
$$\mathbb{E}[q_i \mid \mu_i^K > q^*] = \mathbb{E}[\mu_i^K \mid \mu_i^K > q^*].$$

By proposition 2, $\mu_i^U \sim N(\mu, \sigma_U^2)$ and $\mu_i^K \sim N(\mu, \sigma_K^2)$, with $\sigma_U < \sigma_K$. Hence the conditions of proposition 10 are satisfied, and the result follows. □

Proposition 6.

$$\mu_i^{KA} \sim N\left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2\right),$$
$$\mu_i^{KF} \sim N\left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2\right),$$
$$\mu_i^{UA} \sim N\left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2\right),$$
$$\mu_i^{UF} \sim N\left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2\right).$$

Proof. Since μ_i^{KA} and μ_i^{KF} are simply μ_i^K shifted by a constant (see proposition 5) they follow the same distribution as μ_i^K except that its mean is shifted by the same constant. Similarly μ_i^{UA} and μ_i^{UF} are just μ_i^U shifted by a constant. So the results follow from proposition 2. □

For notational convenience, I introduce q^{KA}, q^{KF}, q^{UA}, and q^{UF}, defined by

$$q^{KA} = q^* + \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \qquad q^{KF} = q^* - \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2},$$
$$q^{UA} = q^* + \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \qquad q^{UF} = q^* - \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}.$$

Theorem 7. *If $\varepsilon > 0$, $\delta > 0$, and $\sigma_{rv}^2 > 0$,*

$$\Pr\left(\mu_i^{KA} > q^*\right) < \Pr\left(\mu_i^{KF} > q^*\right) \text{ and } \Pr\left(\mu_i^{UA} > q^*\right) < \Pr\left(\mu_i^{UF} > q^*\right).$$

Proof. For the first inequality, note that

$$\Pr\left(\mu_i^{KA} > q^*\right) = 1 - \Phi\left(\frac{q^{KA} - \mu}{\sigma_K}\right) < 1 - \Phi\left(\frac{q^{KF} - \mu}{\sigma_K}\right) = \Pr\left(\mu_i^{KF} > q^*\right).$$

The equalities follow from the distributions of the posterior means established in proposition 6. The inequality follows from the fact that Φ is strictly increasing in its argument. By the same reasoning,

$$\Pr\left(\mu_i^{UA} > q^*\right) = 1 - \Phi\left(\frac{q^{UA} - \mu}{\sigma_U}\right) < 1 - \Phi\left(\frac{q^{UF} - \mu}{\sigma_U}\right) = \Pr\left(\mu_i^{UF} > q^*\right).$$

□

Lemma 11.

$$\begin{aligned}\Pr(A_i) &= p_{KA}\left(1-\Phi\left(\frac{q^{KA}-\mu}{\sigma_K}\right)\right)+p_{KF}\left(1-\Phi\left(\frac{q^{KF}-\mu}{\sigma_K}\right)\right)\\ &\quad+p_{UA}\left(1-\Phi\left(\frac{q^{UA}-\mu}{\sigma_U}\right)\right)+p_{UF}\left(1-\Phi\left(\frac{q^{UF}-\mu}{\sigma_U}\right)\right).\\ \mathbb{E}[q_i \mid A_i] &= \mu+\frac{\sigma_K}{\Pr(A_i)}\left(p_{KA}\phi\left(\frac{q^{KA}-\mu}{\sigma_K}\right)+p_{KF}\phi\left(\frac{q^{KF}-\mu}{\sigma_K}\right)\right)\\ &\quad+\frac{\sigma_U}{\Pr(A_i)}\left(p_{UA}\phi\left(\frac{q^{UA}-\mu}{\sigma_U}\right)+p_{UF}\phi\left(\frac{q^{UF}-\mu}{\sigma_U}\right)\right).\end{aligned}$$

Proof. The expression for $\Pr(A_i)$ follows immediately from the distributions of the posterior means established in proposition 6.

To get an expression for $\mathbb{E}[q_i \mid A_i]$, consider first the average quality of scientist i's paper given that it is accepted and given that scientist i is in the group of scientists known to the editor that the editor is biased against. This average quality is

$$\begin{aligned}\mathbb{E}\left[q_i \mid \mu_i^{KA} > q^*\right] &= \mathbb{E}\left[q_i \mid \mu_i^{K} > q^{KA}\right] = \mathbb{E}\left[\mu_i^{K} \mid \mu_i^{K} > q^{KA}\right]\\ &= \mu+\sigma_K R\left(\frac{q^{KA}-\mu}{\sigma_K}\right),\end{aligned}$$

where the first equality simply rewrites the inequality $\mu_i^{KA} > q^*$ in a more convenient form, the second equality uses lemma 8, and the third equality uses equation 1. Similarly,

$$\mathbb{E}\left[q_i \mid \mu_i^{KF} > q^*\right] = \mu + \sigma_K R\left(\frac{q^{KF}-\mu}{\sigma_K}\right),$$
$$\mathbb{E}\left[q_i \mid \mu_i^{UA} > q^*\right] = \mu + \sigma_U R\left(\frac{q^{UA}-\mu}{\sigma_U}\right),$$
$$\mathbb{E}\left[q_i \mid \mu_i^{UF} > q^*\right] = \mu + \sigma_U R\left(\frac{q^{UF}-\mu}{\sigma_U}\right).$$

The average quality of accepted papers $\mathbb{E}[q_i \mid A_i]$ is a weighted sum of these expectations. The weights are given by the proportion of accepted papers that are written by a scientist in that particular group. For example, authors known to the editor that she is biased against form a $p_{KA}\Pr(\mu_i^{KA} > q^*)/\Pr(A_i)$ proportion of accepted papers. Hence

$$\begin{aligned}
\mathbb{E}\left[q_i \mid A_i\right] = {} & \frac{1}{\Pr\left(A_i\right)} p_{KA} \Pr\left(\mu_i^{KA} > q^*\right) \mathbb{E}\left[q_i \mid \mu_i^{KA} > q^*\right] \\
& + \frac{1}{\Pr\left(A_i\right)} p_{KF} \Pr\left(\mu_i^{KF} > q^*\right) \mathbb{E}\left[q_i \mid \mu_i^{KF} > q^*\right] \\
& + \frac{1}{\Pr\left(A_i\right)} p_{UA} \Pr\left(\mu_i^{UA} > q^*\right) \mathbb{E}\left[q_i \mid \mu_i^{UA} > q^*\right] \\
& + \frac{1}{\Pr\left(A_i\right)} p_{UF} \Pr\left(\mu_i^{UF} > q^*\right) \mathbb{E}\left[q_i \mid \mu_i^{UF} > q^*\right] \\
= {} & \mu + \frac{\sigma_K}{\Pr\left(A_i\right)} \left(p_{KA}\phi\left(\frac{q^{KA}-\mu}{\sigma_K}\right) + p_{KF}\phi\left(\frac{q^{KF}-\mu}{\sigma_K}\right)\right) \\
& + \frac{\sigma_U}{\Pr\left(A_i\right)} \left(p_{UA}\phi\left(\frac{q^{UA}-\mu}{\sigma_U}\right) + p_{UF}\phi\left(\frac{q^{UF}-\mu}{\sigma_U}\right)\right). \qquad \square
\end{aligned}$$

References

Charles D. Bailey, Dana R. Hermanson, and Timothy J. Louwers. An examination of the peer review process in accounting journals. *Jour-*

nal of Accounting Education, 26(2):55–72, 2008a. ISSN 0748-5751. doi: 10.1016/j.jaccedu.2008.04.001. URL http://www.sciencedirect.com/science/article/pii/S0748575108000201.

Charles D. Bailey, Dana R. Hermanson, and James G. Tompkins. The peer review process in finance journals. *Journal of Financial Education*, 34: 1–27, 2008b. ISSN 0093-3961. URL http://www.jstor.org/stable/41948838.

Damien Besancenot, Kim V. Huynh, and Joao R. Faria. Search and research: the influence of editorial boards on journals' quality. *Theory and Decision*, 73(4):687–702, 2012. ISSN 0040-5833. doi: 10.1007/s11238-012-9314-7. URL http://dx.doi.org/10.1007/s11238-012-9314-7.

Liam Kofi Bright. Against candidate quality. Manuscript, 2015. URL https://www.academia.edu/11673059/Against_Candidate_Quality.

Liam Kofi Bright, Daniel Malinsky, and Morgan Thompson. Causally interpreting intersectionality theory. *Philosophy of Science*, 83(1):60–81, 2016. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/684173.

Justin Bruner and Cailin O'Connor. Power, bargaining, and collaboration. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*. Oxford University Press, Oxford, forthcoming. URL http://philpapers.org/rec/BRUPBA-2.

Amber E. Budden, Tom Tregenza, Lonnie W. Aarssen, Julia Koricheva, Roosa Leimu, and Christopher J. Lortie. Double-blind review favours increased representation of female authors. *Trends in Ecology & Evolution*, 23(1):4–6, 2008. ISSN 0169-5347. doi: 10.1016/j.tree.2007.07.008. URL http://www.sciencedirect.com/science/article/pii/S0169534707002704.

Kenneth E. Clark. *America's Psychologists: A Survey of a Growing Profession*. American Psychological Association, Washington, 1957.

Jonathan R. Cole and Stephen Cole. Measuring the quality of sociological research: Problems in the use of the "Science Citation Index". *The American Sociologist*, 6(1):23–29, 1971. ISSN 00031232. URL `http://www.jstor.org/stable/27701705`.

Stephen Cole and Jonathan R. Cole. Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review*, 32(3):377–390, 1967. ISSN 00031224. URL `http://www.jstor.org/stable/2091085`.

Stephen Cole and Jonathan R. Cole. Visibility and the structural bases of awareness of scientific research. *American Sociological Review*, 33(3): 397–413, 1968. ISSN 00031224. URL `http://www.jstor.org/stable/2091914`.

Patricia Hill Collins and Valerie Chepp. Intersectionality. In Georgina Waylen, Karen Celis, Johanna Kantola, and S. Laurel Weldon, editors, *The Oxford Handbook of Gender and Politics*, chapter 2, pages 57–87. Oxford University Press, Oxford, 2013. ISBN 0199751455.

Rick Crandall. Editorial responsibilities in manuscript review. *Behavioral and Brain Sciences*, 5:207–208, Jun 1982. ISSN 1469-1825. doi: 10.1017/S0140525X00011316. URL `http://journals.cambridge.org/article_S0140525X00011316`.

Diana Crane. The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist*, 2(4):195–201, 1967. ISSN 00031232. URL `http://www.jstor.org/stable/27701277`.

Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, New Jersey, 2004.

Kenny Easwaran. Probabilistic proofs and transferability. *Philosophia Mathematica*, 17(3):341–362, 2009. doi: 10.1093/philmat/nkn032. URL http://philmat.oxfordjournals.org/content/17/3/341.abstract.

Glenn Ellison. Evolving standards for academic publishing: A q-r theory. *Journal of Political Economy*, 110(5):994–1034, 2002. ISSN 00223808. URL http://www.jstor.org/stable/10.1086/341871.

João Ricardo Faria. The game academics play: Editors versus authors. *Bulletin of Economic Research*, 57(1):1–12, 2005. ISSN 1467-8586. doi: 10.1111/j.1467-8586.2005.00212.x. URL http://dx.doi.org/10.1111/j.1467-8586.2005.00212.x.

Miranda Fricker. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, Oxford, 2007.

Robert D. Gordon. Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941. ISSN 00034851. URL http://www.jstor.org/stable/2235868.

Remco Heesen. Academic superstars: Competent or lucky? *Synthese*, forthcoming. doi: 10.1007/s11229-016-1146-5. URL http://dx.doi.org/10.1007/s11229-016-1146-5.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. John Wiley & Sons, New York, second edition, 1994.

Philip Kitcher. *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford University Press, Oxford, 1993. ISBN 0195046285.

David N. Laband. Publishing favoritism: A critique of department rankings based on quantitative publishing performance. *Southern Economic Journal*, 52(2):510–515, 1985. ISSN 00384038. URL http://www.jstor.org/stable/1059636.

David N. Laband and Michael J. Piette. Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors. *Journal of Political Economy*, 102(1):194–203, 1994. ISSN 00223808. URL http://www.jstor.org/stable/2138799.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. ISSN 1532-2890. doi: 10.1002/asi.22784. URL http://dx.doi.org/10.1002/asi.22784.

D. Lindsey. Using citation counts as a measure of quality in science: Measuring what's measurable rather than what's valid. *Scientometrics*, 15 (3–4):189–203, 1989. ISSN 0138-9130. doi: 10.1007/BF02017198. URL http://dx.doi.org/10.1007/BF02017198.

Christopher D. Mackie. *Canonizing Economic Theory: How Theories and Ideas Are Selected in Economics*. M. E. Sharpe, New York, 1998. ISBN 9780765602848.

Conor Mayo-Wilson, Kevin J. S. Zollman, and David Danks. The independence thesis: When individual and social epistemology diverge. *Philosophy of Science*, 78(4):653–677, 2011. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/661777.

Marshall H. Medoff. Editorial favoritism in economics? *Southern Economic Journal*, 70(2):425–434, 2003. ISSN 00384038. URL http://www.jstor.org/stable/3648979.

Robert K. Merton. A note on science and democracy. *Journal of Legal and Political Sociology*, 1(1–2):115–126, 1942. Reprinted in Merton (1973, chapter 13).

Robert K. Merton. *The Sociology of Science: Theoretical and Empirical Investigations*. The University of Chicago Press, Chicago, 1973. ISBN 0226520919.

Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41): 16474–16479, 2012. doi: 10.1073/pnas.1211286109. URL `http://www.pnas.org/content/109/41/16474.abstract`.

Michael J. Piette and Kevin L. Ross. A study of the publication of scholarly output in economics journals. *Eastern Economic Journal*, 18(4):429–436, 1992. ISSN 00945056. URL `http://www.jstor.org/stable/40325474`.

Jennifer Saul. Implicit bias, stereotype threat, and women in philosophy. In Katrina Hutchison and Fiona Jenkins, editors, *Women in Philosophy: What Needs to Change?*, chapter 2, pages 39–60. Oxford University Press, Oxford, 2013.

Daniel L. Sherrell, Joseph F. Hair, Jr., and Mitch Griffin. Marketing academicians' perceptions of ethical research and publishing behavior. *Journal of the Academy of Marketing Science*, 17(4):315–324, 1989. ISSN 0092-0703. doi: 10.1007/BF02726642. URL `http://dx.doi.org/10.1007/BF02726642`.

Kenneth J. Smith and Robert F. Dombrowski. An examination of the relationship between author-editor connections and subsequent citations of auditing research articles. *Journal of Accounting Education*, 16 (3–4):497–506, 1998. ISSN 0748-5751. doi: 10.1016/S0748-5751(98)

00019-0. URL http://www.sciencedirect.com/science/article/pii/S0748575198000190.

Rhea E. Steinpreis, Katie A. Anders, and Dawn Ritzke. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, 41(7–8):509–528, 1999. ISSN 0360-0025. doi: 10.1023/A:1018839203698. URL http://dx.doi.org/10.1023/A:1018839203698.

Brandon D. Stewart and B. Keith Payne. Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, 34(10):1332–1345, 2008. doi: 10.1177/0146167208321269. URL http://psp.sagepub.com/content/34/10/1332.abstract.

Michael Strevens. The role of the priority rule in science. *The Journal of Philosophy*, 100(2):55–79, 2003. ISSN 0022362X. URL http://www.jstor.org/stable/3655792.

Eric Luis Uhlmann and Geoffrey L. Cohen. "I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, 104(2):207–223, 2007. ISSN 0749-5978. doi: 10.1016/j.obhdp.2007.07.001. URL http://www.sciencedirect.com/science/article/pii/S0749597807000611.

Virginia Valian. *Why So Slow? The Advancement of Women*. MIT Press, Cambridge, 1999. ISBN 9780262720311.

Christine Wennerås and Agnes Wold. Nepotism and sexism in peer-review. *Nature*, 387(6631):341–343, May 1997. ISSN 0028-0836. doi: 10.1038/387341a0. URL http://dx.doi.org/10.1038/387341a0.

Alan J. Ziobrowski and Karen M. Gibler. Factors academic real estate authors consider when choosing where to submit a manuscript for pub-

lication. *Journal of Real Estate Practice and Education*, 3(1):43–54, 2000. ISSN 1521-4842. URL http://ares.metapress.com/content/1762151051KM2227.

Kevin J. S. Zollman. Optimal publishing strategies. *Episteme*, 6:185–199, Jun 2009. ISSN 1750-0117. doi: 10.3366/E174236000900063X. URL http://journals.cambridge.org/article_S1742360000001283.

Strategies of Explanatory Abstraction in Molecular Systems Biology[†]

Nicholaos Jones[‡]

Abstract

I consider three explanatory strategies from recent systems biology that are driven by mathematics as much as mechanistic detail. Analysis of differential equations drives the first strategy; topological analysis of network motifs drives the second; mathematical theorems from control engineering drive the third. I also distinguish three abstraction types: aggregations, which simplify by condensing information; generalizations, which simplify by generalizing information; and structurations, which simplify by contextualizing information. Using a common explanandum as reference point—namely, the robust perfect adaptation of chemotaxis in *Escherichia coli*—I argue that each strategy invokes a different combination of abstraction types and that each targets its abstractions to different mechanistic details.

1 Introductory Remarks

The currently dominant paradigm for understanding explanation in biology puts mechanism at center stage (Nicholson 2012; Levy 2013). Leading accounts of mechanistic explanation, while differing in the particulars of their analysis of *mechanism*, agree that mechanistic explanations explain by alluding to mechanisms or models thereof (Machamer, Darden, Craver 2000; Bechtel and Abrahamsen 2005).

There is a small publishing industry devoted to discerning the scope of mechanistic explanation in scientific practice. Some claim to identify biological explanations that do not allude to mechanisms (Wouters 2007; Huneman 2010; Rice 2015). Fans of mechanistic explanation tend to resist making scope concessions, preferring instead to accommodate the putative explanations as mechanistic despite initial appearances, to broaden the scope of mechanistic explanation or the analysis of *mechanism*, or else to

[†] Draft. For symposium on *Integrating Explanatory Strategies Across the Life Sciences* at the 2016 meeting of Philosophy of Science Association, Atlanta, GA. I thank audiences at Mississippi State University, the Alabama Philosophical Society, and the Society for Philosophy of Science in Practice for comments on earlier drafts.

[‡] Department of Philosophy, University of Alabama in Huntsville, Huntsville AL 35899, nick.jones@uah.edu

deny that the putative explanations are explanations at all (Craver 2006; Bechtel and Abrahamsen 2010; Brigandt 2013; Levy and Bechtel 2013).

I set aside questions about what qualifies as an explanation as well as questions about whether only mechanisms—or models thereof—carry explanatory power. I focus, instead, on *explanatory strategies*, understood as patterns of reasoning directed toward providing explanations. I consider three explanatory strategies from recent systems biology that are driven by mathematics as much as, if not more than, mechanistic detail. Analysis of differential equations drives the first strategy; topological analysis of network motifs drives the second; mathematical theorems from control engineering drive the third.

Systems biologists use these strategies to supplement the explanatory power of traditional molecular mechanisms (see Brigandt et al *forthcoming*). My aim is to identify how the strategies differ from each other, rather than how they differ from standard mechanistic explanations or what might unify them in those differences (for which see Green and Jones 2016). Doing so helps with understanding relations among the strategies, their tactics for integrating mechanistic detail, and explanatory affordances of their mathematical elements.

They key to my analysis is a distinction among three abstraction types: aggregations, which simplify by condensing information; generalizations, which simplify by generalizing information; and structurations, which simplify by contextualizing information. Using a common explanandum as reference point—namely, the robust perfect adaptation of chemotaxis in *Escherichia coli* (Barkai and Leibler 1997; Ma et al 2009; Yi et al 2000)—I argue that each strategy invokes a different combination of abstraction types and that each targets its abstractions to different mechanistic details. I begin with the typology of abstraction.

2 Abstraction Typology

I am interested in abstractions as representational rather than metaphysical. Abstractions, as I understand them, are ontologically innocent, so that characterizing features of representations as abstractions over some parts of reality carries no implication that features correspond to abstract objects (see also Cartwright 1989, 353-354; Levy and Bechtel 2013, 243). So, for example, representing the relation between a person, a hotel, and a date range as a reservation does not entail that some abstract object, a *reservation*, exists; nor does representing the motions of an object's constituents as the motion of the object's center of mass entail that some abstract object, a *center of mass*, exists.

Levy and Bechtel characterize a representation as abstract insofar as a more concrete representation is possible (2013, 242). Brigandt and colleagues suggest that biologists use abstractions to "elucidate system-level patterns of organization that may not be visible at the level of molecular details" (*forthcoming*). I concur. I understand abstractions as representing only some of the many elements—objects, relations, parameters—associated with their targets, thereby making apparent patterns obscured by more detailed representations. I add to these insights that biologists produce (at least) three types of abstraction.

Following Ordorica, I call the first *aggregation* (2015, 163-164). An aggregation represents some relationship among multiple elements of a representational target as a higher-level object, or multiple elements of the target as a single, composite object. (See Figure 1a.) Paradigm cases of aggregations include representations of person-hotel-date relations as *reservations*; of costs of services and costs of goods as *costs*; and of the motions of an object's parts as the *motion of a center of mass* (from Ordorica 2015, 164). Aggregations abstract from plurality to individual, ignoring differences among many in order to make salient some integrated unity among the elements of a representational target. They thereby simplify representations by condensing information about representational targets.

Following Pincock, I call the second abstraction type *generalization* (2015, 864). A generalization represents some element of a representational target as a class of elements, where potential instances of the class might include elements not present in the target. (See Figure 1b.) For example, because the class of solution measures includes all soap-bubble-like surfaces, such as the cellular froth surrounding radiolarian protozoa, representing a soap-bubble surface as a "solution measure" is a generalization (Pincock 2015, 864). Generalizations abstract from an instance to a class thereof, ignoring differences between instances of the class in order to make salient some more general unity. They thereby simplify representations by generalizing from information about representational targets.

I call the third abstraction type *structuration*. A structuration represents some element of a representational target as a position in a structure, such that potential occupants of the position might include elements not present in the target. (See Figure 1c.) I follow Haslanger in understanding structures as "complex entities with parts whose behavior is constrained by their relation to other parts" (2016, 118). Paradigm cases of structurations include representating Barack Obama as President of the United States of America, or representing AIneias as son of Anchises and Aphrodite. Structurations abstract to a position in a structure, from an occupant of the position, ignoring intrinsic features of the occupant unrelated to its position in order to make salient the

occupant's role relative to occupants of other positions in the same structure. They thereby simplify by contextualizing information about representational targets.

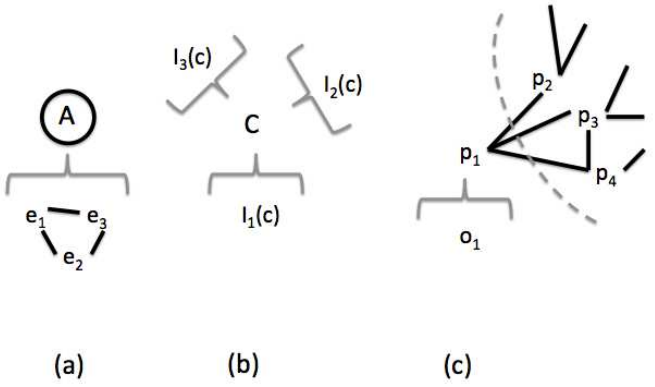


Figure 1: Visualizing Abstraction Types. (a) Aggregation A represents elements e1, e2, and e3 (and relations therein) as a single object. (b) Generalization C represents $I_1(c)$ as a class, instances of which also include $I_2(c)$ and $I_3(c)$. (c) Structuration p_1 represents element o_1 as a position in larger structure that also includes p_2, p_3, and p_4.

I understand aggregations as distinct from both generalizations and structurations, by virtue of being many-to-one, rather than one-to-one, simplifications. I also understand being a generalization as insufficient for being a structuration. For representations of positions carry information about functional relationships between their occupants and other positions in the same structure; but representations of classes do not. Finally, insofar as classes are sets, I understand being a structuration as insufficient for being a generalization. For, sometimes, representing target elements as classes carries some information about intrinsic features of those elements apart from their functional relations to elements occupying other positions in the same structure; but representing target elements as positions in structures never carries such information.

3 Robust Perfect Adaptation of *E.coli* Chemotaxis

My central claim is that different explanatory strategies from recent systems biology differ from each other, at least in part, by virtue of appealing to different abstraction types. I support this claim by considering a case in which multiple strategies target the same explanandum. Doing so minimizes confounds that confuse differences due to the nature of each explanatory strategy with differences due to the nature of each

explanatory target. I focus on a particular explanandum known as robust perfect adaptation of bacterial chemotaxis, following others who consider this a paradigmatic target for non-mechanistic explanation (Braillard 2010; Brigandt, Green, and O'Malley *forthcoming*; Matthiessen *forthcoming*).

3.1 Explanadum Context

Escherichia coli (*E.coli*) is popular model organism in biological research. It is very sensitive to small chemical changes over a very large range of background concentrations. It also has a simple and well-understood signal transduction network (Wadhams and Armitage 2004).

E.coli manages two kinds of motion (Berg 2003). It *runs* by rotating its flagellar motor counterclockwise. This aligns all of its flagella into a synchronized bundle, resulting in movement in a straight line for about 1 second. *E.coli* also *tumbles* by rotating its flagellar motor clockwise. This breaks flagellar alignment, and the asynchronized flagella produce stationary changes of direction lasting for about 0.1 second. *E.coli* are randomly reoriented after each tumble. Moreover, while these tumbles occur with regular frequency, *E.coli* with higher concentrations of CheR protein tumble more frequently (Spudich and Kochland 1975).

E.coli's motion in a uniform external environment resembles a random walk. *E.coli* has no ability to control or select its direction of motion, and its straight runs are subject to Brownian motion because of eddies. However, in the presence of a chemical attractant—amino acids such as serine or aspartic acid, or sugars such as maltose or glucose— *E.coli taxis* toward the attractant. This taxi behavior involves less frequent tumbles, leading to longer runs and so gradual motion toward the attractant. (There is an opposite behavior for repellants such as metal ions or leucine.)

The biomolecular mechanism for *E.coli* chemotaxis is well-understood. When an environmental attractant attaches to a receptor, the receptor lowers the activity of the CheW-CheA protein complex. Less activity from this complex reduces the rate of CheY phosphorylation, which results in less phorphorylated CheY diffusing to the flagella. Because CheY induces clockwise rotation of the flagellar motor, the outcome is less frequent tumbling.

3.2 Explanadum Question

Alon and colleagues have experimental verification that, in the presence of a chemical attractant mixed uniformly into the environment at a constant concentration, *E.coli* chemotaxis *perfectly adaptive* (Alon et al 2009). After a brief period of decreased tumbling frequency, the frequency of *E.coli* tumbles increases toward and returns to the

exact frequency prior to the introduction of the attractant. The effect of the attractant, accordingly, becomes entirely forgotten despite its continuing presence.

The biomolecular mechanism for the adaptiveness of chemotaxis for *E.coli* is also well-understood. Some time after a new attractant has been detected by receptors, the lower activity of the CheW-CheA complex induces less CheB activity. This reduces the rate for removing methyl groups from the CheW-CheA complex and, together with continual methylation of the CheR receptor, CheW-CheA methylation increases. More methylation means more CheW-CheA activity, which in turn induces more CheY phosphorylation. This eventually results in more phosphorylated CheY diffusing to the flagellar motor, which increases clockwise motor rotation and thereby raises tumbling frequency.

Alon and colleagues have further experimental verification that this perfectly adaptive chemotaxis of *E.coli* is *robust* across ranges of CheR concentrations 0.5 to 50 times higher than concentration levels in "wild type" *E.coli* (Alon et al 2009). (By contrast, *E.coli*'s adaptation time—the time to return to 50% of its pre-stimulus tumbling frequency—is not robust to different CheR concentrations, because more CheR entails longer adaptation times.) This is the explanandum of interest: why is the perfect adaptation of *E.coli* chemotaxis, in the presence of a well-distributed chemical attractant, robust to CheR protein concentrations?

There are (at least) three strategies for answering this question in recent systems biology literature. (For a fourth, see Kollman et al 2005.) I consider each in turn, first sketching the general strategy and then making explicit the abstractions at work.

4 Distinguishing Explanatory Strategies through Abstraction Types

4.1 Dynamical Modeling

I call the first strategy *dynamical modeling*. This strategy begins by constructing a chemotaxis network for *E.coli*. This network represents the mechanism for *E.coli* chemotaxis, including specific biochemical details about when and how relevant proteins affect each other. (See Figurer 2.) For example, Barkai and Leibler (1997) construct a model according to which, among many other specifics, CheB demethylates only the active form of the CheW-CheA complex and CheR works only at saturation.

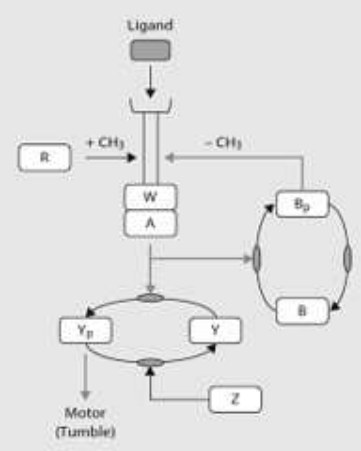


Figure 2. Mechanistic network for *E.coli* chemotaxis (Rao and Ordal 2009).

The dynamical modeling strategy proceeds by constructing a dynamical model—typically a set of differential equations—from the network (see Jones and Wolkenhauer 2012). One then demonstrates, via mathematical proof or simulation, that this model predicts perfect adaptation in the presence of a well-distributed chemical attractant for CheR concentration values varying over several orders of magnitude. (Raerinne 2013 calls this *sensitivity analysis*.) The demonstration supports the inference that *E.coli* chemotaxis exhibits robust perfect adaptation *because of its biochemical specifics*.

Bechtel and Abrahamsen (2010) call the product of this strategy a *dynamical mechanistic explanation*. I set aside the issue of whether the dynamical modeling strategy produces explanations. But I endorse Bechtel and Abrahamsen's insight that the dynamical modeling strategy produces accounts that are mechanistic, by virtue of depending upon mechanistic details, as well as dynamical, by virtue of analyzing mathematical models built upon those details. For example, Barkai and Leibler's (1997) mathematical analysis is relevant to *E.coli* chemotaxis only insofar as their network details are relevant; and analysis of the network apart from the model cannot produce an inference about the *robustness* of *E.coli*'s perfectly adaptive chemotaxis.

Let's treat the dynamical model driving this explanatory strategy as an initial baseline for evaluating the number and severity of abstraction in various explanatory strategies. The model is abstract in various ways. But we shall treat it as a recipient of further abstractions, in the way a vehicle receives freight. Just as we can determine the weight of the freight indirectly by subtracting the gross weight of vehicle and freight from the "tare weight" (the weight of vehicle alone), we shall determine abstraction variety and

severity/extent for models driving other explanatory strategies by "subtracting" their total abstraction variety and severity from the "tare" abstraction.

4.2 Topological Analysis

I call the second explanatory strategy *topological analysis*. This strategy begins by identifying all possible minimal adaptation networks capable of predicting robust perfect adaptation for *E.coli* chemotaxis. These networks, like the networks for dynamical modeling, represent mechanisms for *E.coli* chemotaxis. Yet, unlike the networks for dynamical modeling, these networks are minimal: they contain the fewest possible nodes and links that suffice for robustly perfectly adaptive chemotaxis. The procedure for identifying all possible minimal networks of this sort is brute computational search. It turns out that there are exactly three, each of which has exactly three nodes and no more than three links (Ma et al 2009).

The topological analysis strategy proceeds by identifying a chemotaxis network known to predict robust perfect adaptation. This strategy thereby relies upon the dynamical modeling strategy, but only for mathematical results. The biochemical details of the chosen chemotaxis network turn out to be largely irrelevant, because the topological analysis strategy proceeds by demonstrating that a *reduced form* of the chosen network is topological equivalent to one of the minimal adaptation models. Reduced forms for mechanistic networks functional equivalents for node groups, group nodes or equivalents into modules, and ignore links within modules in favor of links between modules.

Consider, for example, one of the three minimal adaptation networks Ma and colleagues (2009) discover for *E.coli* chemotaxis. (See Figure 3.) The network has an input activating node A, A inhibiting being activated by B, A also activating C, and C activating some output. Ma and colleagues show that Barkai and Leibler's (1997) model for *E.coli* chemotaxis reduces to this minimal network. Barkai and Leibler have an input and CheR activating, and CheB inhibiting, receptors; these receptors activating the CheW-CheA complex; the complex activating CheB and CheY; and CheY activating some output. Ma and colleagues reconceptualize Barkai and Leibler's network into one where the input activates a *receptor complex*; this complex activates CheY, which activates the output; the complex also activates CheB, which inhibits a *methylation level* also activated by CheR.; and this methylation level activates the receptor complex. Then, in a second reconceptualization that produces one of their minimal adaptation networks, they group the receptor complex and CheB into module A, group CheR and the methylation level into module B, and rename CheY module C.

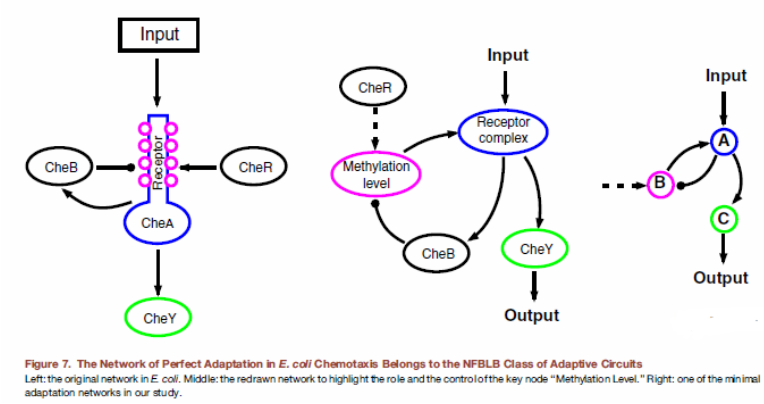


Figure 7. The Network of Perfect Adaptation in *E. coli* Chemotaxis Belongs to the NFBLB Class of Adaptive Circuits
Left: the original network in *E. coli*. Middle: the redrawn network to highlight the role and the control of the key node "Methylation Level." Right: one of the minimal adaptation networks in our study.

Figure 3: Network topology for *E.coli* chemotaxis (Ma et al 2009).

The topological analysis strategy infers, from the topological equivalence between a minimal adaptation network and the reduced form of a network known to predict robust perfect adaptation for chemotaxis, that *E.coli* chemotaxis exhibits robust perfect adaptation *because of the topology of its chemotaxis network*. Huneman (2010) calls the product of this strategy a *topological explanation*. Regardless of whether analyses such as Ma and colleagues's are explanatory, they are topological by virtue of demonstrating some consequence about the topological properties of a network. This means that, even if the mechanistic details of *E.coli*'s chemotaxis network were different, and even if the biochemical specifics of the network chosen for reduction were different, the product of the topological analysis strategy would remain the same provided that the alternative networks preserve topological equivalence with the originals (see also Jones 2014).

The topological model driving this second explanatory strategy is more abstract than the dynamical model driving our initial ("tare") strategy. The topological model contains more aggregations. For example, it represents CheY and CheZ as "the motor rotation group;" it represents CheA and CheW as "the receptor complex;" and it represents the receptor complex and CheB as "the phosphorylation group." The topological model also contains more structurations. For example, it represents the phosphorylation group as "A" and the motor rotation group as "C." These representations abstract entirely from any intrinsic marks that might distinguish instances of "A" from instances of "C," relying instead upon extrinsic relations to distinguish the nodes from each other. So, for example, "A" but not "C" inhibits "B," "A" activates "C," and so on.

4.3 Organizational Design

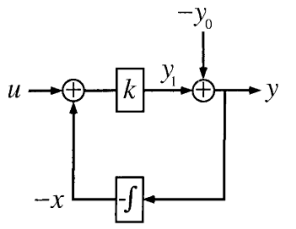
I call the third explanatory strategy *organization design*. This strategy begins with a proof to the effect that systems exhibit robust perfect adaptation if and only if they

satisfy the characteristic equation for Integral Feedback Control (IFC). The proof is purely mathematical, well-known from control engineering theory in contexts involving mechanical systems that exhibit IFC such as thermostats. I am not aware of a complete and published version of this proof, but Yi and colleagues (2000) provide a sketch with relevant details. The organizational design strategy proceeds by inferring that *E.coli* chemotaxis exhibits robust perfect adaptation if and only if it satisfies the characteristic equation for IFC, and further inferring that *E.coli* chemotaxis exhibits robust perfect adaptation *because it satisfies the characteristic equation for IFC*. (For better explanatory details regarding this specific case, Braillard 2010; Green and Jones 2016.)

The organizational design strategy invokes neither mechanistic specifics about the chemotaxis network for *E.coli* nor topological details about the structure of that network. The strategy takes the explanandum phenomenon as given, using a mathematical equivalence result to identify a principle both necessary and sufficient for the phenomenon. The strategy thereby has affinities with explanatory strategies that appeal to organizing principles (Green and Wolkenhauer 2013) and design principles (Green 2015).

For simplicity, let's "reset" our abstraction "tare" to the topological model, because the model driving the organizational design strategy—call it the design model—is abstract in all the ways the topological model is abstract and more besides. The simplification thereby focuses attention on ways in which the design model differs from the topological model—and, by extension, from the initial dynamical model.

Compared to the topological model, the design model contains more aggregations. For example, the design model represents CheY phosphorylation and CheB activation as "*k*-box output." This aggregation is, at the same time, a generalization and a structuration. For example, "*k*-box output" is a class, with instances biological as well as mechanical. The standard example of a mechanical instance is heater activation in a thermostat. The *k*-box representation is also a structuration, akin to the "A", "B," and "C" representations from the topological model. For the *k*-box represents whatever has such-and-such input and output (a position in a structure). (See Figure 4.)



$$\begin{aligned} \dot{x} &= y & y(t) &\to 0 \text{ as } t \to \infty \\ y &= y_1 - y_0 & \textit{iff} & \\ &= k(u-x) - y_0 & k &> 0 \end{aligned}$$

Fig. 2. A block diagram of integral feedback control. The variable u is the input for a process with gain k. The difference between the actual output y_1 and the steady-state output y_0 represents the normalized output or error, y. Integral control arises through the feedback loop in which the time integral of y, x, is fed back into the system. As a result, we have $\dot{x} = y$ and $y = 0$ at steady-state for all u. In the Barkai–Leibler model of the bacterial chemotaxis signaling system, the chemoattractant is the input, receptor activity is the output, and $-x$ approximates the methylation level of the receptors.

Figure 4 Organizational design for bacterial chemotaxis…and thermostats (Yi et al 2000).

The topological model is more abstract than the dynamical model, by virtue of containing various abstractions over protein identities. The design model, in turn, is more abstract than the topological and dynamical models, by virtue of also containing various abstractions over protein interactions. We can, therefore, arrange the various explanatory strategies along a continuum of abstraction type and severity. The dynamical modeling strategy, as our baseline, occupies the "low" end of our continuum. Next is topological analysis, which involves aggregations of and structurations from protein identities (or aggregations thereof). Then there is organizational design, which also involves aggregation of protein interactions as well as generalization and structuration of protein identities (or aggregations thereof).

5 Confirming the Analysis

I consider the foregoing to establish that each explanatory strategy invokes a different combination of abstraction types and that each targets its abstractions to different mechanistic details. Whether this result generalizes beyond my chosen case study awaits future research. There is some reason to expect an affirmative result. For if dynamical, topological, and design explanatory strategies differ as I claim—specifically, along dimensions of number and severity of generalizations and structurations—then we should expect the *more abstract* strategies to have *wider scope*. For the more general models likely have more instances, and the more structural models likely have more position occupants.

We find confirmation of this prediction for the case of robust perfect adaptation of *Bacillus subtilis* (*B.subtilis*) chemotaxis. Details of the organization design strategy for explaining why *E.coli* chemotaxis exhibits robust perfect adaptation *also* apply for explaining why *B.subtilis* chemotaxis exhibits robust perfect adaptation. But details of the corresponding dynamical mechanistic strategy do not. The organization design strategy, as we know, involves more generalization and structuration than the dynamical mechanistic strategy. This confirms our prediction.

Allow me to be brief with the details. Rao and Ordal (2007) develop a dynamic mechanistic explanation for the perfect robustness of chemotaxis for *B.subtilis*. Their explanatory strategy follows the same pattern as Barkai and Leibler's in the case of *E.coli*. But details differ. For example, according to Barkai and Leibler's model, CheB in *E.coli* demethylates only active receptor complexes; according to Rao and Ordal, CheB in *B.subtilis* demethylates inactive ones too. Again, according to Barkai and Leibler's model, without CheY *E.coli* runs but does not tumble; according to Rao and Ordal, without CheY *B.subtilis* tumbles but does not run. One more: according to Barkai and Leibler's model, *E.coli* without CheB cannot run; according to Rao and Ordal, *B.subtilis* without CheB can run. See Figure 5.

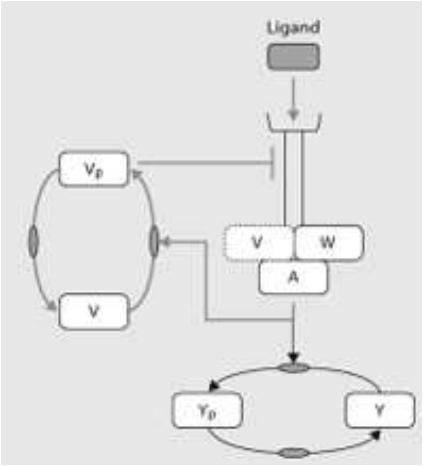


Figure 5: Chemotaxis network for *B.subtilis* (Rao and Ordal 2009).

So Barkai and Leibler's dynamical mechanistic explanation does not apply for the case of *B.subtilis*. But Yi and colleague's organizational design strategy does. For *B.subtilis*, like *E.coli*, exhibits robust perfect adaptation for chemotaxis if and only if it satisfies the characteristic equation for integral feedback control.

6 Toward Abstractive Mechanistic Explanation and its Affordances

Systems biological strategies for explaining the robust perfect adaptation of bacterial chemotaxis (in *E.coli*, *B.subtilis*, etc) apply mathematical techniques to network models. Dynamical, topological, and design strategies apply different techniques to explain the same phenomenon. Each explanatory strategy, moreover, applies its mathematical techniques to network models that embody different kinds and severities of these abstractions such as aggregations, generalizations, structurations. These abstraction types, accordingly, help to explain how these systems biological explanatory strategies differ from each other.

These abstraction types also provide a foundation for unifying various explanatory strategies from systems biology under the banner of mechanistic explanation. Let's consider well known kinds of mechanistic explanation as *standard*. Let's also follow Bechtel and Abrahamsen (2010) by considering *dynamical* mechanistic explanation as a mathematized species of standard mechanistic explanation.

Then let an *abstract network* be any network representation obtained by aggregating, generalizing, or structurating mechanistic details of the sort familiar in standard mechanistic explanation. Also let an *abstractive* mechanistic explanation be any explanation driven by applying mathematical techniques to an abstract network. See Figure 6.

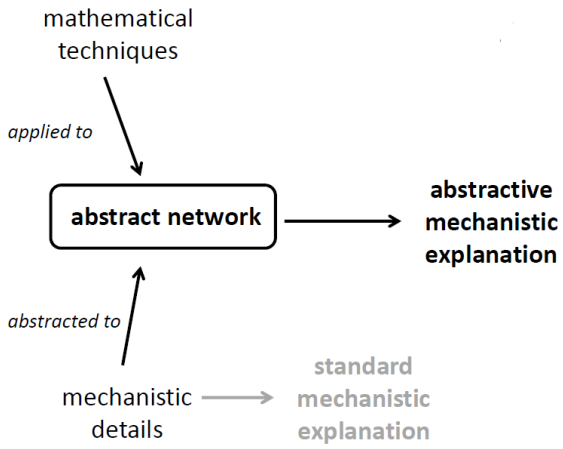


Figure 6. Relating standard and abstractive mechanistic explanation.

Then topological and organizational design explanatory strategies are mechanistic strategies—albeit abstractive ones. Topological explanations apply topological analysis

to aggregated and generalized mechanism networks. Organizational design explanations apply control systems engineering to aggregated, generalized, and structurated mechanism networks.

Both kinds of explanation are mechanistic, by virtue of being grounded upon mechanistic details. But both also provide explanatory affordances unavailable through standard mechanistic explanations, by virtue of being abstract. For example, by virtue of using generalizations, topological explanations should have a greater scope than their standard mechanistic counterparts. By virtue of using generalizations and structurations, organizational design explanations should have still greater scope.

That these abstractive mechanistic strategies use novel mathematical techniques is a side effect of their using novel abstractions (in comparison with standard mechanistic explanations and their dynamical cousins). These techniques, of course, support more general conclusions, with wider scope, than the kind of differential equation analysis available for dynamical mechanistic explanations. But the techniques do not explain why the strategies have broader scope.

References

U.Alon, M.G.Surette, N.Barkai, and S.Leibler, “Robustness in Bacterial Chemotaxis,” *Nature* 397 (2009), 168-171.

N.Barkai and S. Leibler. “Robustness in simple biochemical networks,” *Nature* 387 (1997), 913-917.

W.Bechtel and A.Abrahamsen, "Explanation: A Mechanistic Alternative," *Studies in History and Philosophy of Biological and Biomedical Science* 36 (2005), 421-441.

W.Bechtel and A.Abrahamsen, "Dynamic Mechanistic Explanation: Computational Modeling of Circadian Rhythms as an Exemplar for Cognitive Science," *Studies in History and Philosophy of Science* 41 (2010), 321-333.

H.C.Berg, *E.coli in Motion* (Springer, 2003).

P.A. Braillard, “Systems Biology and the Mechanistic Framework,” *History and Philosophy of the Life Sciences* 32.1 (2010), 43-62.

I.Brigandt, "Systems Biology and the Integration of Mechanistic Explanation and Mathematical Explanation," *Studies in History and Philosophy of Biological and Biomedical Sciences* 44 (2013), 477-492.

I.Brigandt, S.Green, and M.O’Malley, “Systems Biology and Mechanistic Explanation,” in S. Glennan and P. Illari (eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (*forthcoming*).

N.Cartwright, “Capacities and Abstraction,” in P. Kitcher and W. Salmon (eds.), *Scientific Explanation* (University of Minnesota Press, 1989), 349-356.

C.F.Craver, "When Mechanistic Models Explain," *Synthese* 153 (2006), 355-376.

S.Green, “Revisiting Generality in Biology: Systems Biology and the Quest for Design Principles,” *Biology and Philosophy* 30.5 (2015), 629-652.

S.Green and N.Jones, “Constraint-Based Reasoning for Search and Explanation: Strategies for Understanding Variation and Patterns in Biology,” *dialectica* 70.3 (2006), 343-374.

S.Green and O.Wolkenhauer, “Tracing Organizing Principles: Learning from the History of Systems Biology,” *History and Philosophy of the Life Sciences* 35 (2013), 553-576.

S. Haslanger, "What is a (Social) Structural Explanation?" *Philosophical Studies* 173 (2016), 113-130.

P.Huneman, "Topological Explanation and Robustness in Biological Sciences," *Synthese* 177 (2010), 213-245.

N.Jones, “Bowtie Structures, Pathway Diagrams, and Topological Explanation,” *Erkenntnis* 79.5 (2014), 1135-1155.

N.Jones and O.Wolkenhauer, “Diagrams as Locality Aids for Search and Explanation in Molecular Cell Biology,” *Biology and Philosophy* 27 (2012), 1135-1155.

M.Kollman, L.Løvdok, K.Bartholome, J.Timmer, and V.Sourjik, "Design Principles of a Bacterial Signalling Network," *Nature Letters* 438.24 (2005), 504-507.

A.Levy, “Three New Kinds of Mechanism,” *Biology and Philosophy* 28.1 (2013), 99-114.

A.Levy and W.Bechtel, “Abstraction and the Organization of Mechanisms,” *Philosophy of Science* 80 (2013), 241-261.

W.Ma, A.Trusina, H.El-Samad, W.A.Lim, and C.Tang, “Defining Network Topologies that Can Achieve Biochemical Adaptation,” *Cell* 138 (2009), 760-773.

P.Machamer, Lindley Darden, and C.F.Craver, "Thinking about Mechanisms," *Philosophy of Science* 67 (2000), 1-25.

D.Matthiessen, “Mechanistic Explanation in Systems Biology: Cellular Networks,” *British Journal for Philosophy of Science* (*forthcoming*).

D.J.Nicholson, “The Concept of Mechanism in Biology,” *Studies in History and Philosophy of Biologica and Biomedical Sciences* 43.1 (2012), 152-163.

S.A.G.Ordorica, “The Explanatory Role of Abstraction Processes in Models: The Case of Aggregations,” *Studies in History and Philosophy of Science* (2015), 161-167.

C. Pincock, “Abstract Explanations in Science,” *British Journal for the Philosophy of Science* 66 (2015), 857-878

J.Raerinne, “Robustness and Sensitivity of Biological Models,” *Philosophical Studies* 166 (2013), 285-303.

C.V.Rao and G.W.Ordal, “The Molecular Basis of Excitation and Adaptation during Chemotaxic Sensory Transduction in Bacteria,” in M. Collin and R. Schuch (eds.), *Bacterial Sensing and Signaling* (Karger: Basil, Switzerland, 2009), 33-64.

C.Rice, "Moving Beyond Causes: Optimality Models and Scientific Explanation," *Nous* 49 (2015), 589-615.

J.L.Spudich and D.E.Kochland, “Non-Genetic Individuality: Chance in the Single Cell,” *Nature* 262 (1976), 467-471.

G.H.Wadhams and J.P.Armitage, “Making Sense of It All: Bacterial Chemotaxis,” *Nature Reviews: Molecular Cell Biology* 5 (2004), 1024-1037.

A.G.Wouters, "Design Explanation: Determining the Constraints on What Can Be Alive," *Erkenntnis* 67 (2007), 65-80.

T.-M.Yi, Y.Huang, M.I.Simon, and J.Doyle, “Robust Perfect Adaptation in Bacterial Chemotaxis through Integral Feedback Control,” *PNAS* 97 (2000), 4649-4653.

How the Diachronic Theoretical Virtues Make an Epistemic Difference

Mike Keas • Professor of the History and Philosophy of Science • The College at Southwestern

Abstract. Among the virtues of good theories are those appropriately labeled diachronic: durability, fruitfulness, and applicability—the last of which is insufficiently recognized. Diachronic theoretical virtues *cannot* be instantiated in the original construction of a theory; subsequent development is required. By contrast, one *can* assess the degree to which a theory exhibits the following nine non-diachronic theoretical virtues in a theory's original construction: evidential accuracy, causal adequacy, explanatory depth, internal consistency, internal coherence, universal coherence, beauty, simplicity, and unification. The distinction between diachronic and non-diachronic virtues is important for understanding the role and epistemic standing of each theoretical virtue.

Keywords. Theoretical virtues, durability, fruitfulness, prediction, and science-technology relations.

1. Introduction. Theoretical virtues are the traits of a theory that show it is probably true or worth accepting. Although the identification, characterization, classification, and epistemic standing of theory virtues are debated by philosophers and by participants in specific theoretical disputes, many scholars agree that these virtues help us to infer which rival theory is the best explanation (Lipton 2004). The most widely accepted theories across the disciplines usually exhibit many of the same theoretical virtues listed below. Each virtue class contains at least three virtues that sequentially follow a repeating pattern of progressive disclosure or expansion. In another forthcoming essay (Keas 2017) I argue for this new systematization of the theoretical virtues. In the present essay I focus on the diachronic class of virtues in contrast with the non-diachronic virtues. One can assess the degree to which a theory exhibits the non-diachronic virtues from the time a theory is initially framed. However, no theory, in its original construction, can instantiate

the diachronic virtues: durability, fruitfulness, or applicability. These virtues are instantiable only as a theory is later refined or applied.

Evidential virtues
1. Evidential accuracy: A theory (T) fits the empirical evidence well (regardless of causal claims).
2. Causal adequacy: T's causal factors plausibly produce the effects (evidence) in need of explanation.
3. Explanatory depth: T excels in causal history depth or in other depth measures such as the range of counterfactual questions that its law-like generalizations answer regarding the item being explained.

Coherential virtues
4. Internal consistency: T's components are related to each other logically.
5. Internal coherence: T's components are coordinated into an intuitively plausible whole; T lacks ad hoc hypotheses—theoretical components merely tacked on to solve isolated problems.
6. Universal coherence: T sits well with (or is not obviously contrary to) other warranted beliefs.

Aesthetic virtues
7. Beauty: T evokes aesthetic pleasure in properly functioning and sufficiently informed persons.
8. Simplicity: T explains the *same facts* as rivals, but with *less* theoretical content.
9. Unification: T explains *more kinds of facts* than rivals with the *same* amount of theoretical content.

Diachronic virtues
10. Durability: T has survived testing by successful prediction or plausible accommodation of new data.
11. Fruitfulness: T has generated additional discovery by means such as successful novel prediction, unification, and non ad hoc theoretical elaboration.
12. Applicability: T has guided strategic action or control, such as in science-based technology.

We will survey the first nine virtues only to the brief extent needed to recognize how one can assess the degree to which a theory exhibits these theoretical virtues in its original construction. This will, by contrast, enable us to appreciate the unique temporal character of the diachronic theoretical virtues.

2. Non-Diachronic Theoretical Virtues. We begin with the first three virtues. *Evidential accuracy*, which is how well a theory fits the relevant data, can be assessed from the theory's original construction. Often a theory will also, from its inception, specify *causally adequate* mechanisms to produce the phenomena in question. Such is not necessarily the case, as Alfred Wegener's theory of continental drift illustrates. His theory enjoyed considerable evidential accuracy despite its lack of a plausible cause to move the continents. *Explanatory depth* is also instantiated in a theory's initial formulation if, for example, the

theory answers a large range of counterfactual questions about a kind of phenomenon using the resources of its law-like generalizations.

The remaining six non-diachronic theoretical virtues likewise can be exhibited in the initial formation of a theory. A theory may be constructed in a logical manner so as to produce *internal consistency*. Beyond that, the theoretical components might be well coordinated into an intuitively plausible whole (avoiding ad hoc hypotheses), thus generating the theoretical virtue of *internal coherence*. If the theory sits well with (or is not obviously contrary to) other warranted beliefs, then it possesses the virtue of *universal coherence*. A new theory might even evoke aesthetic pleasure in the minds of experts, which constitutes theoretical *beauty*. The closely related virtues of simplicity and unification also might be instantiated in the initial formation of a theory: explaining the same facts as rival theories but with less theoretical content (*simplicity*), and explaining more kinds of facts than rivals with the same amount of theoretical content (*unification*).

Much more could be said about the first nine virtues outlined above (Keas 2017), but this is sufficient to recognize them as a group of theoretical virtues that can, in principle, be instantiated in a theory's original formation. This common trait remains characteristic of these virtues even (largely) under the disparate accounts found in the literature of how to characterize each virtue. Let us now explore the chief diachronic theoretical virtues in contrast to the non-diachronic virtues.

3. Diachronic Theoretical Virtues. Durability, fruitfulness, and applicability, which I recognize as the chief diachronic theoretical virtues, can only be instantiated as a theory is cultivated *after* its origin. This necessarily extended temporal dimension of the diachronic virtues is, arguably, of considerable epistemic importance. But even if one endorses the arguments that discount the epistemic significance of this temporal component (Mayo 2014), one still should acknowledge a group of virtues that (unlike the other the-

oretical virtues) can only be instantiated in a theory *after* its initial formulation. Time is of their essence in a manner that goes beyond the trivial truth that all human endeavor is temporal. McMullin (2014) has lead the way in articulating the epistemic significance of two of the three main diachronic virtues: durability and fruitfulness (I recognize McMullin's third diachronic virtue of "consilience" as a mode of fruitfulness). Applicability, largely overlooked as a theory virtue, is another important member of this diachronic category, as I shall demonstrate.

3.1. Durability. Durability, a virtue term McMullin (2014) recommended, refers to the favorable epistemic condition of a theory that has survived testing by successful prediction or by plausible accommodation of new unanticipated data (or both). Popular or long-lived theories are not necessarily durable in the epistemic sense in view here. Equating durability with popularity or tradition is fallacious. While testability is a pragmatically admirable trait of a theory, it is not an intrinsic epistemic characteristic of a theory; many testable theories have failed too many tests to be acceptable. Steel (2010, 18) notes that the "more precise and informative a theory's empirical predictions are, the greater its testability." The more testable a theory is, the more durable it would prove itself to be if it passes the tests. A theory that scores low in testability has little potential to exhibit durability.

Despite the leading role of predictive success in many areas of science, it is less prominent in some reputable scientific theories that are, nevertheless, well endowed with other virtues. Successful prediction is very frequently part of explaining "how things work," but less routine in explaining "how things originated"—as in theories about the history of the cosmos, earth, and life (Cleland 2011, but Winther 2009 argues otherwise). Successful historical theories typically enjoy other forms of durability, most notably a track record of plausible accommodation of new data that, although not predicted, came to light after the theory's origin. The durability of a theory suffers if one or more of its predictions are disconfirmed

or when theorists respond to disconfirming evidence by modifying the theory with ad hoc hypotheses—theoretical components merely "tacked on" to solve isolated problems. Although initially a theory may exhibit a high degree of evidential accuracy (or any other of the first nine virtues in my systematization), it is impossible for a newborn theory to instantiate the virtue of durability—this *takes time* in a sense not required by the non-diachronic virtues. A similar necessary temporal dimension characterizes fruitfulness.

3.2. Fruitfulness. Fruitfulness, also known as fertility or fecundity, is another diachronic theoretical virtue. A theory is fruitful if, over time, it generates additional discovery by means such as successful novel prediction, unification, and non ad hoc theoretical elaboration. While durability is about conservation (a theory passing tests to survive), fruitfulness is about innovation (a theory stimulating further discovery). When a prediction formulated in the context of a theory's construction is later verified, this successful predictive outcome increases the virtue of durability in that theory. By contrast, a *novel* prediction is one that was not conceived in conjunction with a theory's construction, but that nevertheless follows reasonably from it. When such a novel prediction is confirmed by observation, a theory exhibits more fruitfulness.

The closely related diachronic character of durability and fruitfulness is well illustrated in the discovery of the first two planets beyond Saturn. Soon after Friedrich William Herschel unexpectedly discovered Uranus in 1781, astronomers noted that its observed motion strayed from what contemporary Newtonian mechanics predicted of such a planet. However, given the overall theoretically virtuous status of Newtonian physics up through that time (including its durability due to its success in testing), most astronomers expected a forthcoming way to make Uranus compliant with established theory. Even rejecting the anomalous data as "inaccurate" seemed reasonable early on. By the 1830s, however, the possibility of a perturbing planet beyond Uranus became a more reasonable and popular speculation, despite the ab-

sence of a precise novel prediction of where to find such a planet. By this time many astronomers were modestly confident in the accumulated data of Uranus' positions in the sky.

This brings us to the celebrated successful novel prediction of 1845-1846. Based principally on Newtonian physics and the well-known irregularities in Uranus' motion, two astronomers independently predicted where another unknown perturbing planet (later called Neptune) was likely located. Le Verrier's estimate of the planet's location was the most accurate (correct within one degree), as confirmed by a German astronomer on September 23, 1846. The (*fruitful*) novel prediction of Neptune was born within the context of a *durable* Newtonian orbital mechanics research tradition and the unexpected discovery of Uranus with its anomalous motions. The sensational success of this novel prediction (the discovery of Neptune) also rendered Uranus a Newtonian-compliant planet—thus further vindicating earlier provisional toleration of Uranus' anomalies, a toleration that had been justified by yet earlier Newtonian durability and fruitfulness.

Smith's (2010; 2014) landmark study of gravity theory from Newton to the present further illuminates the durability and fruitfulness of this research tradition, and it includes the case histories of Uranus and Neptune. Smith was surprised that the principal kind of question being tested was not "Do the calculated motions [e.g., of Uranus] agree with the observed motions?" Rather it was: "Can robust physical sources compatible with Newtonian theory be found for each clear, systematic discrepancy between the calculated and the observed motions?" Neptune (as novelly predicted) turned out to be such a robust physical source. However scientists failed over a half century to find a robust (detectable) physical source for the Newtonian-defying behavior of Mercury—a tiny anomaly in the precession of its perihelion. But this failure, which Einstein solved by way of theory replacement, does not completely diminish the enduring epistemic significance of two centuries of Newtonian durability and fruitfulness, as Hanson (1962) inaccurately suggested. Smith notes: "All the other discrepancies ended up revealing some detail of our plane-

tary system, the least subtle of which was Neptune, that theretofore had not been taken into account in the calculations" (2010, 552).

Such serial Newtonian problem solving became (almost always) ever more empirically constrained in a spiral of upward progress. For example, Uranus' temporarily Newtonian-defying behavior "would have been masked if the significantly larger gravitational effects of Saturn on Uranus had not been included in the calculation first." Smith explains further:

> So, the discovery of Neptune provided evidence not only for Newton's theory, but also for the specific aspects of Saturn that entered into calculating its effects on Uranus, for these were no less presupposed in the anomaly that emerged than Newton's theory was. The point generalizes. Each time a discrepancy emerges and a robust physical source for it is found, that source is incorporated into the new calculations, and the process is repeated, typically with still smaller discrepancies emerging that were often theretofore masked in the calculations. So, what was being tested each time when a new discrepancy emerged and a physical source for it was being sought was not only Newtonian theory, but also all the previously identified details that make a difference and the differences they were said to make without which the further systematic discrepancy would not have emerged. (2010, 552-53)

On display is an interlocking of durability (passing tests to survive) and fruitfulness (stimulating further discovery) that is supportive of scientific realism. "This shows that increasingly strong evidence was accruing to Newtonian theory over the first two hundred years of orbital research based on it," Smith concludes. This point (with some qualification) extends even to Einstein's theoretical innovation that was partly justified by the unruly perihelion of Mercury. Einstein's achievement was, to some degree, a continuation of this same progressive spiral, as Smith deftly explains:

> As is well known, Einstein required Newtonian gravitation to hold in an asymptotic limit as he developed his new theory of gravity—specifically in a static, weak-field limit. That he did so was just as well because the 43 arc-seconds per century anomaly in the perihelion of Mercury that was initially the sole evidence for his theory presupposes Newtonian gravity…. As a matter of historical fact, all of the details singled out as making detectable differences during the two centuries of prior research carried over intact into post-Einstein orbital mechanics. *Save for some qualifications concerning levels of precision, the same details are still making the same differences as before.*... So, Newtonian theory must still have some sort of claim to being knowledge. (2010, 556-57)

Smith's continuity-of-knowledge claim invites comment. While much of the metaphysics associated with Newtonian theory has been repudiated, we nevertheless see an impressive degree of fruitful scientific continuity from Newtonian to modern physics (at least in the particular ways that Smith documents). In sum, Newtonian orbital mechanics enjoyed increasingly impressive interlocking durability and fruitfulness over multiple centuries, and its approximate legitimacy (not counting discarded Newtonian metaphysics) remains similarly well-grounded today under the revisionary umbrella of modern physics.

Though some philosophers have argued to the contrary (Collins 1994; Harker 2008), many scientists and philosophers think that predictive success—especially novel predictive success—is a stronger indicator of likely approximate truth than a theory's accommodation of data (Douglas and Magnus 2013). According to my systematization (which illuminates but does not settle this thorny issue), data accommodation refers to a theory's initial instantiation of the evidential virtues (evidential accuracy, causal adequacy, and explanatory depth), and a theory's subsequent instantiation of certain diachronic virtues, namely non-predictive durability (plausibly making sense of new unanticipated data) and non-predictive fruitfulness (especially non ad hoc theoretical elaboration that makes sense of new unanticipated data).

3.2.1 Unification as a Mode of Fruitfulness. Fruitful theory elaboration, whether by means of successful novel prediction or non ad hoc theoretical elaboration that makes sense of unanticipated evidence, often also makes sense of *new kinds* of data, and thus is additionally recognized as increasing a theory's unification. Earlier we encountered unification as a non-diachronic (aesthetic) theoretical virtue. The diachronic increase of unification differs somewhat from its non-diachronic cousin. The historian and philosopher of science William Whewell (1794–1866) called diachronic unification "consilience." When a theory explains a new domain of facts in a surprising way, then it is fruitful in a consilient manner. McMullin writes in this regard:

> A good theory will often display remarkable powers of unification, making different classes of phenomena "leap together" over the course of time. Domains previously thought to be disparate now become one, the textbook example, of course, being Maxwell's unification of magnetism, electricity, and light. Examples abound in recent science, a particularly striking one being the development of the plate-tectonic model in geology. Assuming that this unifying power manifests itself over time, it testifies to the epistemic resources of the original theory and hence to that theory's having been more than mere accommodation. (2014, 505)

McMullin contrasts diachronic unification with its non-diachronic counterpart: "If the unification was achieved by the original theory, however, the virtue involved would no longer be diachronic." Instead, it would count (in my systematization) as an aesthetic theoretical virtue that I simply call "unification," and that Lipton calls "variety" (and yet others call "broad scope"). Lipton favors the assumption that such "heterogeneous evidence provides more support than the same amount of very similar evidence" (Lipton 2004, 168). Despite my own inclination to accept Lipton's point, I recognize this as a somewhat debatable assumption about the epistemic significance of an aesthetic property. However, when unification increases

over time, especially by means of surprising convergences, then unification is less likely the result of the idiosyncratic aesthetic predispositions and clever accommodating skills of a theorist during theory formation. Thus fruitful diachronic unification has greater confirmatory power than a theory's initial degree of aesthetic unification.

3.2.2 The Role of Prediction in the Diachronic Virtues. Drawing from Douglas' work on the relationship of prediction to inferring the best explanation, I argue that predictive success (in the first two diachronic virtues explored above) extends the epistemic work of many non-diachronic theoretical virtues such as causal adequacy, explanatory depth, beauty, simplicity, and unification. These latter theory traits, which she collectively labels as "explanatory,"

> appeal to us, not just because we are aesthetically driven creatures but because such virtues help us to use the explanation to think and, in particular, to think our way through to new predictions, new tests, new rigors for our beautiful explanation. (2009, 460)

Douglas also notes:

> Predictions are valuable because they force us (when followed through) to test our theories, because they have the potential to expand our knowledge into new realms and because they hold out the possibility (if successful) of gaining some measure of control over natural processes. (2009, 455)

Transposing Douglas' insights into my taxonomic terms, predictions are valuable because they figure into all three of the major diachronic virtues: durability (testing theories successfully), fruitfulness (expanding "our knowledge into new realms"), and applicability (which includes "gaining some measure of control over natural processes"). Moreover, the operation of prediction ("saying before" at least in a logical if not

temporal sense) in these three theoretical virtues further supports my classification of them as diachronic. Lets us now explore the last major diachronic virtue of applicability.

3.3. Applicability. Applicability refers to when a theory is used to guide successful action (e.g., prepare for a natural disaster) or to enhance technological control (e.g., genetic engineering). High degrees of the virtue of applicability obtain when a theory that is used to guide such action or control provides more effective outcomes than what is possible in the absence of the theory. Successful scientific theories constitute *knowledge* of the world (knowing *that*), not *control* over the world (which is mainly knowing *how*) for practical (non-theoretical) purposes. In this regard Strevens (2008, 3) notes: "If science provides anything of intrinsic value, it is explanation. Prediction and control are useful … but when science is pursued as an end rather than as a means, it is for the sake of understanding." But even after the intrinsic good of a theoretically virtuous explanation is in hand, one of several possible additional confirmatory diachronic (predictive or controlling) virtues might be acquired by a theory, including applicability. In such cases a good theory just gets better—even more confidence in its probable truth is justified.

Although scientific experiments use technological control, they do so to test scientific theories—so the main function is still to understand nature, not to control it. However, especially in the case of theories supported by experimentally verified prediction, such foreknowledge and laboratory control might be exploited to achieve practical aims such as device fabrication or medical intervention. But in any case, one cannot *apply* scientific knowledge until *after* one first *obtains* it. This necessary time lapse makes applicability diachronic.

To obtain scientific knowledge we search for a theory that (initially) exhibits many of the non-diachronic theoretical virtues. Subsequent work aimed at theory testing and elaboration might produce the additionally confirming presence of the diachronic virtues of durability and fruitfulness. At some point in

this dance of virtue-driven theory assessment and refinement, sufficient confidence in a particular theory might spur attempts to apply it as the basis for a new or improved technology. If the derived science-based technology actually works, then the "applied theory" has acquired the additional theoretical virtue of applicability. Because this requires additional time after initial theory formation, the diachronic classification of applicability is appropriate.

Although the application of scientific theories constitutes one aspect of technology, most of technology involves the empirical discovery of "know how" knowledge without crucially presupposing or immediately applying any particular scientific theory. Indeed, the relation between science and technology is not a simple one-way linear affair (Radder 2009; Douglas 2014). But this "emancipation" of technology from subordination to science, accomplished by historians and philosophers of technology between 1960 and 1990 (Houkes 2009, 310), should not obscure the epistemic significance of instances of technological innovation made possible, in part, by *applied* scientific theory.

This point is in harmony with the so-called demise of the "pure vs. applied science" dichotomy. Understanding and controlling nature are closely related, as our study of the diachronic theoretical virtues, including applicability, indicates. Douglas (2014, 62) surfaces some of the subtlety of this argument when, on the one hand, she proclaims: "With the pure vs. applied distinction removed, scientific progress can be defined in terms of the increased capacity to predict, control, manipulate, and intervene in various contexts." But then, on the other hand, in a footnote she recoils partially: "To be clear, while I think this is a useful rubric for scientific progress, it is not a remotely sufficient account for how one should assess scientific theories." Other (non-diachronic) theoretical virtues that are complementary to, but less weighty epistemically than, prediction and control also play important roles in theory assessment, she suggests. Consideration of the nine major non-diachronic theoretical virtues systematized in Sections 1 and 2 drives this point home.

How exactly is applicability a diachronic theory trait that is *epistemic* (helping to indicate likely truth) in view of the obvious *pragmatic* orientation of technological application? Agazzi observes that some technological projects "are designed or projected in advance, as the concrete application of knowledge provided by a given science or set of sciences" (Agazzi 2014, 308). If a project of this kind actually works as predicted, then this reinforces our confidence in the theory base that helped guide such action in the world. Agazzi further notes:

> The predictions 'contained' in the project actually are the predictions made by the scientific theories which have permitted the proposal of the complex *noema* that constitutes the project, and contains not only prescriptions as to the way of realising the structure of the machine but also as to its functioning. This functioning is something that happens; it is a state of affairs that constitutes a confirmation of the theories used in projecting the machine. (309)

Although Agazzi's scientific realism overstates the epistemic reach of applicability, it is helpful nonetheless as a corrective to other philosophical errors:

> A mature science is a science that has given rise to a significant technology. This means, for example, that we can provisionally admit certain theories that are 'empirically adequate,' without admitting their truth as van Fraassen says, until we have significant predictions confirming them. This fact (especially in conjunction with other 'virtues' discussed in the literature) already justifies attributing truth and ontological reference to them, but the existence of technological applications is the last decisive step that assures that they have been able to adequately treat those aspects of reality they intended to treat. These last words are very important. They underline the fact that technological success does not eliminate the partial or limited scope of scientific theories. The fact that we can use classical mechanics in creating many machines or for sending rockets into space certainly means that this mechanics is true of its

> objects and therefore 'tells a true story' about certain aspects of reality. This can also be expressed by saying that this theory is partially true of reality, but only if we mean that it does not speak about the totality of the attributes of reality, and that, consequently, it can speak properly only of such referents that possess these attributes. In other words, it is not correct to say that this mechanics is true regarding the whole of reality because other aspects of reality exist that must be accounted for by means of other theories which, in turn, can be used as a basis for different technologies. (310-11)

To nuance Agazzi's insightful but somewhat inflated epistemic role for applicability, we can observe that this theoretical virtue is not commonly operative in certain scientific domains. For example, scientific theories of "how things originated" (history of nature) lead to fewer technological applications than scientific theories of "how things work." Part of the reason for the infrequent applicability of origins theories is the smaller role that experimentally controlled prediction plays in such theorization. For example, much of the data that allows us to reconstruct the *history* of earth's surface is collected by means of passive field observations, rather than by laboratory experiments that make precise predictions and technological control more feasible.

4. Conclusion. The diachronic theoretical virtues possess a temporal dimension that is absent from the other theoretical virtues. They can only be instantiated *after* a theory's initial formulation—when it has had opportunity to be tested, elaborated, and applied. Durability, fruitfulness, and applicability build upon the initial theory assessment process governed by the non-diachronic virtues (the evidential, coherential, and aesthetic theoretical virtues). The cumulative result, when successful, is a mature theory with an even greater probability of being true than an infant theory that has not yet had the opportunity to show whether it will possess the diachronic theoretical virtues (anti-realists are invited to interject their own alternative

to this realist understanding of the theoretical virtues). So, the distinction between diachronic and non-diachronic virtues is important for an adequate account of theory evaluation.

The three major diachronic theoretical virtues are also better understood when they are recognized as related to each other in the following progressive sequence. Durability is instantiated as a theory passes more rigorous tests in a series of encounters with the world, especially by successful prediction and plausible accommodation of new evidence. Fruitfulness discloses a theory's resourcefulness yet further through innovation—stimulating additional discovery by successful novel prediction, unification, non ad hoc theoretical elaboration, and other means. At last, applicability expands the epistemic accountability of a theory into the final frontier: the vast domain of practical action. This virtue is instantiated when a theory helps us to interact with the world successfully, most notably by technological control. Together, these diachronic theoretical virtues provide an ongoing and epistemically intensified means of theory development that complements the non-diachronic virtue assessment process that begins in a theory's original construction.

Applicability, *as a theoretical virtue*, has not received the attention it deserves. Surprisingly, it is absent from every theoretical virtue list I have encountered. My work sketches a way to understand applicability in relation to the other diachronic virtues, and the larger group of non-diachronic virtues. This endeavor promises to illuminate, among other things, discussion of realism vs. anti-realism, science-technology relations, and inference to the best explanation.

References

Agazzi, Evandro. 2014. *Scientific Objectivity and Its Contexts*. Cham: Springer.

Cleland, Carol E. 2011. "Prediction and Explanation in Historical Natural Science." *British Journal for the Philosophy of Science* 62 (3):551-582.

Collins, Robin. 1994. "Against the Epistemic Value of Prediction over Accommodation." *Noûs* 28 (2):210-224.

Douglas, Heather E. 2009. "Reintroducing Prediction to Explanation." *Philosophy of Science* 76 (4):444-463.

——— 2013. "The Value of Cognitive Values." *Philosophy of Science* 80 (5):796-806.

——— 2014. "Pure Science and the Problem of Progress." *Studies in History and Philosophy of Science Part A* 46:55-63.

Douglas, Heather, and P. D. Magnus. 2013. "State of the Field: Why Novel Prediction Matters." *Studies in History and Philosophy of Science Part A* 44 (4):580-589.

Hanson, Norwood Russell. 1962. "Leverrier: The Zenith and Nadir of Newtonian Mechanics." *Isis* 53 (3):359-378.

Harker, David. 2008. "On the Predilections for Predictions." *The British Journal for the Philosophy of Science* 59 (3):429-453.

Houkes, Wybo. 2009. "The Nature of Technological Knowledge." In *Philosophy of Technology and Engineering Sciences*, 309-350. Amsterdam: North-Holland.

Lipton, Peter. 2004. *Inference to the Best Explanation*. 2nd ed. London: Routledge.

Mayo, D. 2014. "Some Surprising Facts About (the Problem of) Surprising Facts." *Studies in History and Philosophy of Science Part A* 45:79-86.

McMullin, Ernan. 2014. "The Virtues of a Good Theory." In *The Routledge Companion to Philosophy of Science*, ed. Martin Curd and Stathis Psillos, 561-571. New York: Routledge.

Radder, Hans. 2009. "Science, Technology and the Science-Technology Relationship." In *Philosophy of Technology and Engineering Sciences*, ed. A. Meijers, 65-91. Amsterdam: North Holland.

Smith, George E. 2010. "Revisiting Accepted Science: The Indispensability of the History of Science." *Monist* 93 (4):545-579.

——— 2014. "Closing the Loop: Testing Newtonian Gravity, Then and Now." In *Newton and Empiricism*, ed. Zvi Biener and Eric Schliesser, 262-351. New York: Oxford University Press.

Steel, Daniel. 2010. "Epistemic Values and the Argument from Inductive Risk." *Philosophy of Science* 77 (1):14-34.

Strevens, Michael. 2008. *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Winther, R. G. (2009). Prediction in selectionist evolutionary theory. *Philosophy of Science, 76*(5), 889-901.

Reconciling axiomatic quantum field theory with cutoff-dependent particle physics

Adam Koberinski[1]

[1]Department of Philosophy, Western University

Abstract

The debate between Fraser and Wallace (2011) over the foundations of quantum field theory (QFT) has spawned increased focus on both the axiomatic and conventional formalisms. The debate has set the tone for future foundational analysis, and has forced philosophers to "pick a side". The two are seen as competing research programs, and the major divide between the two manifests in how each handles renormalization. In this paper I argue that the terms set by the Fraser-Wallace debate are misleading. AQFT and CQFT should be viewed as complementary formalisms that start from the same physical basis. Further, the focus on cutoffs as demarcating the two approaches is also highly misleading. Though their methods differ, both axiomatic and conventional QFT seek to use the same physical principles to explain the same domain of phenomena.

1 Introduction

Foundational invesitgation into quantum field theory (QFT) has emerged as a flourishing enterprise in philosophy of science, thanks largely to work done in axiomatic QFT (AQFT), particularly the C^*-algebraic approach enocoded by the Haag-Kastler axioms (Haag and Kastler 1964). Despite the methodological disconnect with 'conventional' approaches to QFT (CQFT), AQFT has been defended by Fraser (2009) as supplying a firmer foundation from which to conduct philosophical analyses. Though this is one of few explicit defenses of AQFT, the widespread use of algebraic methods in philosophical literature on QFT would lead one to believe that Fraser is merely making explicit the assumptions in her field. Recently, Wallace (2006; 2011) has questioned the focus on AQFT, arguing that CQFT is the better candidate for analysis. Since CQFT is the theory that has been emprically successful—the Standard Model of particle physics is built from CQFTs—and AQFT has yet to reproduce these results, Wallace argues that we should focus analysis on CQFT rather than AQFT. Fraser's (2011) reply has set up what is now known as the Fraser-Wallace debate over the foundations of QFT. The debate has set the tone for future foundational analysis, and seems to force philosophers to "pick a side"—you either work in AQFT or CQFT. The two are seen as competing research programs, and the major divide between the two manifests in how each handles renormalization. AQFT requires strict Poincaré covariance at arbitrarily small length scales, while the renormalization group (RG) methods in CQFT allow for a small-scale cutoff, below which QFTs needn't be well-defined.

In this paper I argue that the terms set by the Fraser-Wallace debate are misleading. One needn't view AQFT and CQFT as rival research programs; in fact, this view is

detrimental to understanding the history and methodology of QFT. AQFT and CQFT should be viewed as complementary formalisms that start from the same physical basis. Further, the focus on cutoffs as demarcating the two approaches is also highly misleading: AQFT can accommodate cutoffs and RG methods, and CQFT does not explicitly require cutoffs. The focus on cutoffs as essential to CQFT could mistakenly be taken to mean that CQFT depends on cutoffs actually *being physical*, in the same way that cutoffs are physical in condensed matter physics (CMP). I will argue that this is not the case: cutoffs needn't be physical in any sense. Even if cutoffs are *physically significant*, that does not entail that the cutoffs are themselves physical. Specifically, RG methods provide no principled grounds for thinking that cutoffs are "real" in the sense of signifying a breakdown of field theories generally. Since Wallace (2011) set the terms of the debate, the bulk of the arguments in this paper will be in reference to that paper. I do not claim that Wallace holds all (or even most) of the views against which I argue; rather, I use his paper to clarify potential misconceptions that could arise from the debate. Renormalization is not central to the physical content of QFT, and the different ways of handling renormalization do not mark AQFT and CQFT as different research programs. We should instead view the formalisms as complementary: though their methods differ, both seek to use the same physical principles to explain the same domain of phenomena.

2 Renormalization and the relationship between AQFT and CQFT

Wallace (2011) emphasizes the ineliminable dependence on cutoffs in CQFT, along with the success of RG methods for providing a physical motivation for cutoffs, as the wedge which drives AQFT and CQFT apart. For Wallace, AQFT cannot deal with physical cutoffs. Since RG methods have physically legitimized cutoffs, AQFT and CQFT have differing physical content and must therefore be considered a different research program (2011, Sec. 2). I disagree with this characterization on two fronts. First, AQFT has the resources to incorporate RG methods when needed. Though typical axioms make no metion of scaling behaviour, even the most rigid of axiomatic approaches—algebraic QFT as codified in the Haag-Kastler axioms—can incorporate something like RG flows.[1] Second, the calculational dependence on cutoffs in CQFT may not signal the physical existence of cutoffs.

So, are cutoffs really that problematic for AQFT? Many axiomatic approaches to QFT make no recourse to cutoffs, either explicitly or implicitly. An explicit forbidding of cutoffs would mean that one of the axioms/postulates of the theory claimed that the theory is empirically adequate at all spacetime length scales. Even if any axiomatization contained such an axiom (none do), it would be hard to imagine what sort of work it would do in derivations. Presumably, such a system could be modified to remove the guilty axiom, without spoiling any physically useful theorems. One should therefore not be concerned with an explicit ban on cutoffs in AQFT.

The more interesting case is when cutoffs are implicitly rejected by a particular theory.

[1] See Buchholz and Verch (1995) for an example of scaling algebras playing the role of RG flows.

There are two common assumptions in AQFT that are problematic for handling cutoffs: strongly continuous implementations of Lorentz invariance, and the association of algebras with arbitrarily small open bounded regions of spacetime. Though the latter is not common to all axiomatic QFTs (the Wightman axioms deal directly with quantum fields, rather than algebras), the dominant axiomatization in terms of C^* algebras—the Haag-Kastler axioms—define QFTs in terms of algebras of observables corresponding to open, bounded regions of spacetime.[2] It is implicit that for any open bounded spacetime region, *no matter how small*, one can define an algebra of observables satisfying the other axioms defining QFT. If cutoffs are physical, one might conclude that there should be a principled limit to the size of regions on which we can define algebras corresponding to observables in QFT. If the cutoff scale is physically relevant, and only CQFT predicts its existence, we might be tempted to conclude that the two are different, competing theories. However, there are several possibilities for reconciling AQFT and cutoffs, which I will outline below. These remedies are largely independent of one another, and organized in terms of increasing foundational disagreement with Wallace's view of cutoffs. The "quick fixes" proposed first lead to further conceptual worries, and I therefore endorse the option in Sec. 2.3, which is the biggest departure from taking cutoffs as physical in CQFT. Nevertheless, all the options sketched below are more-or-less viable. Section 2.4 outlines reasons for thinking that *both* AQFT and CQFT suffer the same conceptual challenges if cutoffs *really are physical*.

[2]Since algebraic QFT is prima facie the most problematic, I will deal primarily with algebraic QFT in this paper. The reader can take AQFT to stand for axiomatic QFT or algebraic QFT for the remainder of this paper. The reader should also note that constructive QFT is another important strand of rigorous QFT. Though it is conceptually distinct from AQFT, the two projects often overlap.

2.1 Possibilities for cutoffs in AQFT

Just because we need to associate an algebra with any arbitrary open bounded region of spacetime, we are not therefore compelled to make this algebra interesting. One way that cutoffs could be introduced into AQFT is to specify that regions smaller than some 4-volume Λ are to be uniformly assigned trivial algebras, i.e., algebras containing only multiples of the identity. Such assignments would be consistent with the demand that all open bounded regions of spacetime be assigned an algebra, but it would make the cutoff physically relevant, since no information about local parameters would be contained in regions smaller than Λ.

Though this solution is available, it is admittedly somewhat ad hoc. Even worse, it violates one of the crucial Haag-Kastler axioms: that of weak additivity. The axiom of weak additivity states that, for *every* closed, bounded region $\mathcal{O}$ of Minkowski spacetime $\mathcal{M}$, the C^* norm closure of the algebras $\mathfrak{A}(\mathcal{O}+\alpha)$ for $\alpha \in \mathbb{R}^4$ is just the quasilocal algebra for the whole spacetime, $\mathfrak{A}(\mathcal{M})$.[3] There are two reasons why this is a problem for introducing cutoffs in the way described above. First, we run into the problem that the quasilocal algebra corresponding to the whole of $\mathcal{M}$ can be constructed from *any* algebra corresponding to *any* closed, bounded region $\mathcal{O}$. The norm closure of extensions of a trivial algebra will not produce any interesting algebra as a result, so regions smaller than the cutoff Λ will violate weak additivity. Second, extensions of an arbitrary region $\mathcal{O}$ by some $\alpha < \Lambda$ should not be physical if Minkowski spacetime breaks down at scales below Λ. In the spirit of the first ad hoc axiom modification, weak additivity could be modified to exclude regions $\mathcal{O}_{small} < \Lambda$, and arbitrary extensions $\alpha_{small} < \Lambda$. However,

[3]See Ruetsche (2011), especially chapters 4 and 5 for an introduction to algebraic QFT. For a more comprehensive review of algebraic QFT, see Halvorson and Müger (2007).

there seems to be no principled reason for choosing a specific value of Λ, and one may question the naturalness of such axioms. This makes the solution of simple axiom modification less tempting, and forces us to admit that AQFT—at least in its current guise—is in conflict with approaches to QFT that take cutoffs as physically meaningful, since the basic axioms are currently in direct conflict with the introduction of cutoffs. If we admit that there is currently no room in the formalism of AQFT for cutoffs, are we doomed to take AQFT as (incorrectly) positing its own validity at all energy scales?

2.2 No cutoffs? No problem

If QFT methods are only applicable up to some cutoff energy, and we expect QFT to incorporate this fact, we are saying that a good theory should signal its own demise. The formal necessity of cutoffs in the formalism of CQFT has lead to the idea that our best theories will continue to be an increasing hierarchy of effective field theories. Each field theory requires cutoffs to be implemented at a certain energy scale, and this signals the field theory's domain of applicability. If supplanted by a successor field theory, one expects that the new theory's low energy regime reduces to the old theory, and further that the new theory will itself have a higher energy cutoff. Following this approach, the conventional formalism of field theories would allow us to climb higher and higher up the ladder of energy scales, but we would never reach the top. We would require a theory of a fundamentally different formal type in order to end the ladder of cutoffs. This is presumably the view that Wallace holds, as he claims that if we replace one field theory with another applicable at higher energies, "that field theory in turn will need some kind of short-distance cutoff" (2011, p. 118).

As great as it may be to have a framework in which theories limit their own domain of

applicability, this is certainly not a necessary condition that any good formalism need satisfy. Even if AQFT does not contain cutoffs explicitly, this does not make it at odds with CQFT. Many theories that have been useful in the past do not signal their ultimate demise; on the contrary, most are mathematically well-defined well beyond their domain of applicability. For example, classical theories of fluid dynamics treat fluids as classical continua, and these continua are uniform to arbitrary precision. Classical continuum fluid dynamics is a useful theory, and compatible with classical point mechanics, even though classical point mechanics leads one to believe that the continuum is only an approximation—at some point fluid dynamics must break down. There is nothing within the formalism of fluid mechanics that signals its eventual breakdown; rather, the physical systems we model using classical fluid dynamics, as well as the complementary formalism of classical point particles, give us a physical motivation for the eventual breakdown of the formalism. Deeper theories, such as quantum mechanics, also provide grounds for believing in the limited applicability of both of the complementary classical formalisms. Similarly, we can view AQFT as a complementary picture to the formalism of CQFT. Both formalisms rely on the same general physical principles, though they are implemented in different ways. Though the AQFT formalism does not demarcate its domain of applicability in the form of explicit cutoffs, the necessity of some form of cutoff in CQFT provides reason to believe that the AQFT formalism is only approximately mapping the actual physics. Further, whatever extratheoretical grounds we have for taking cutoffs to be physical—typically in the guise of speculative physics beyond the Standard Model—can inform the scale at which we lose faith in the predictions of *both* the AQFT and CQFT formalisms. When one does not view AQFT and CQFT as rival research programs, the two can work together to provide a deeper

physical understanding of high energy physics, and the role of cutoffs is made clearer.

2.3 Physical significance versus being physical

Are cutoffs really that central? The arguments in the previous section assume that the cutoffs required to generate predictions in CQFT are physical, in the sense that they signal a breakdown of QFT. The fact that perturbative calculations within a particular model diverge when the integrals are unbounded does not entail that field theoretic methodology loses physical significance near these bounds. Undoubtedly we have extratheoretical reasons for supposing that the QFTs making up the Standard Model are not accurate to arbitrary energies—at some point gravity will surely play an important role, to say nothing for possible unknown physics at higher energy scales—but this needn't signify a breakdown of QFTs *in general* beyond a cutoff. Nor is this notion built in to the conceptual apparatus of RG methods, as Wallace claims.[4] It remains entirely possible that a QFT built with more terms in its Lagrangian could describe all relevant physics and be well-defined at all energy scales. In fact, the renormalization group procedure presupposes a theory given in terms of a Lagrangian or Hamiltonian with an arbitrary number of terms. These terms are shown to go to zero in the low energy limit (Wilson and Kogut 1974). We know—using the RG methods to determine the flow of coupling constants—that for non-Abelian gauge theories, interactions become weaker at higher energy scales. Total asymptotic freedom would be one way to eliminate cutoffs at

[4] "Wilsons explanation of the renormalisation procedure relies upon *the failure of the QFT to which it is applied* at very short distances. It is then intriguing to ask how to put on a firm conceptual footing a theory which relies for its mathematical consistency on its own eventual failure". (Wallace 2006, 34, emphasis added) Again, this passage can be read in a way that agrees with the arguments of this section. I am attempting to argue against a naive reading, which takes the failure of *one* QFT (i.e., a single form of interaction, encoded in a particular Lagrangian) to signal the failure of QFT methods in general.

high energies. A successor QFT, such as a grand unified theory or supersymmetry, could therefore unite the strong and electroweak coupling constants, while remaining well-defined to arbitrarily high energies.[5] All that RG methods rely on conceptually is the ability to average out behaviour at high energy scales, and this is compatible with many options for high-energy behaviour. First, our theories could be low-energy approximations that break down at higher energy scales. This could be due to a fundamental granularity or discreteness in the more fundamental theory, or due to the absence of terms in the Lagrangian modelling high energy dynamics. Second, we could have a well-defined high energy dynamics that is unimportant at the energy scales with which we are concerned. In any case, RG methods provide no principled grounds for thinking that cutoffs are "real" in the sense of signifying a breakdown of field theories generally. Unlike the breakdown of classical fluid mechanics—for which we have a more fundamental successor theory (quantum mechanics) providing grounds to reject the continuum as merely an approximation—there is as of yet no (empirically successful) fundamental successor theory for which QFT can be considered a continuum approximation.

One of the major reasons for thinking that cutoffs in QFT mark a regime beyond which the methods of QFT can no longer be applied is the success of RG methods originating from CMP (Wallace 2011, Sec. 1). RG methods were initially developed to investigate long range correlations in materials approaching a phase transition. Long range interactions are those most relevant to global transitions of a material, and so RG

[5]Whether a theory can be made well defined for arbitrarily high energies is a distinct issue from the accuracy of that theory's predictions at high energies. It may turn out that Standard Model QFTs can be extended in a consistent way, but that the high energy predictions turn out to be false. This is the case that is argued in Section 2.2 regarding AQFT.

methods average out the unimportant short range behaviour near a critical point. The apparatus of non-relativistic QFT (i.e., functional integrals using Galilean invariant Lagrangians) is used in CMP as an *approximation* to the discrete atomic (or ionic) physical makeup of bulk systems. Given the the CMP field theories are explicitly constructed as approximations to a known underlying lattice model, we know that the field theoretic methods must break down within CMP. RG flow equations are derived by separating field variables φ into low- and high-momentum components $\varphi = \varphi_{low} + \varphi_{high}$ (where the cutoff from low to high is chosen arbitrarily) and averaging over the high momentum modes. The resulting Lagrangian $\mathcal{L}'(\varphi_{low})$ is then manipulated to fall into the same form as the original Lagrangian $\mathcal{L}(\varphi)$. This process is repeated and generates discrete recursive relations between the rescaled coupling parameters in the $(n+1)$th Lagrangian in terms of the nth one. In the limit where the rescalings are continuous, these become differential equations determining the flow of coupling constants under RG. As the flows are taken to zero frequency—equivalent to the infinite spatial limit—only those parameters relevant to phase transitions will remain in the renormalized Lagrangian. One of the most qualitatively interesting features of successively averaging out short distance (and therefore high energy) degrees of freedom is that, no matter how complicated the initial field dynamics are (encoded as a Lagrangian), only the renormalizable terms will contribute to the low energy dynamics of the theory. This implies that a very broad class of higher energy Lagrangians can "reduce" to the relevant dynamics at lower energy scales.

The success of RG methods in CMP lead to their quick application in QFTs (Wilson 1983)[6], since the relevant formalism is shared between the two disciplines. If we choose

[6]Wilson even forms the QFT/statistical mechanics analogy explicitly, though the source analog in that

to endow the RG methods with similar physical significance in QFT, then we can interpret the high energy cutoffs required as marking the domain at which we expect new physics to occur. The problem is that, because RG flows tell us that our low-energy (effective) QFTs are largely insensitive to the dynamical details at higher energies, they provide little insight our guidance into the high energy physics. Though the path to the successor theory isnt apparent given our current QFTs, the up side is that our best QFTs are protected from the details of our ignorance of high energy dynamics.
Where Wallace might be read to err is in the jump from believing that cutoffs have physical relevance in QFTs to believing that cutoffs *are physical*:

> "This, in essence, is how modern particle physics deals with the renormalization problem: it is taken to presage an ultimate failure of quantum field theory at some short lengthscale, and once the bare existence of that failure is appreciated, the whole of renormalization theory becomes unproblematic, and indeed predictively powerful in its own right" (Wallace 2011, p. 119).[7]

The difference is subtle. Cutoffs can be *physically relevant* in that they signal the breakdown of the *particular* theory or model beyond a certain energy scale, but whether cutoffs themselves *are physical* depends on the precise nature of the breakdown. If the

case is a classical Ising model (Wilson and Kogut 1974). Fraser (2016) has provided an in-depth analysis of the elements of the analogies between QFT and the Ising model, as well as the process of describing RG flow.

[7]Or at least this is a jump he is sometimes guilty of. In other places he is more careful to elaborate on this view, and it appears that he at least appreciates the fact that field theoretic methods may not break down at all (Wallace 2006, pp. 43-4). As mentioned in the introduction, this paper is not a critique of Wallace's view explicitly, but of the misleading way of framing AQFT and CQFT as rivals based on their differing treatments of the arbitrarily small; for this reason I aim to clarify the mistakes in a "naive" reading of Wallace.

breakdown can be remedied by adding new terms in the Lagrangian—effectively changing the particular theory, but retaining the field theoretic framework—then the cutoffs signal new physics, but are not themselves physical. If the breakdown is due to the inapplicability of field theoretic methodology beyond that scale, then the cutoffs are themselves physical.[8] Even if one takes the cutoffs to have physical significance, cutoffs needn't *be physical* in this stronger sense.

One possible reason for thinking that cutoffs are physical is based off of reading too much into the analogy with CMP. We know that field theoretic methods are approximations in bulk matter systems—the atomic theory implies that macroscopic matter is composed of discrete components. The analogy between QFT and CMP is based on the use of the same field theoretic formalism in both disciplines, not on a well-grounded physical similarity.[9] Cutoffs are physical in CMP field theory because field theoretic methods have been introduced as an approximation. Given that discrete quantum mechanics of 10^{23} particles is intractable, we sacrifice (a surprisingly small amount of) precision in order to apply the more soluble methods developed in QFT. But the fact that cutoffs signal the breakdown of field approximations in CMP does not imply that the same is true in QFT. The reasons we treat cutoffs as physical in CMP are absent in QFT; there is no empirically successful theory that claims QFT breaks down due to an underlying discreteness of physics near cutoff scales. Speculative physics may posit some underlying structure for which quantum fields are merely an approximation,

[8]Presumably, the failure of field theoretic methodology in general would require some physical granularity at high energies. This is what I mean by the cutoff being physical and is in direct analogy with the case of non-relativistic QFT in CMP.

[9]Fraser (2016) and Fraser and Koberinski (2016) provide two concrete examples of fruitful formal analogies between QFT and CMP. In the former case, it is the RG flow that is formally analogous, while the latter deals with the formal similarities between spontaneous symmetry breaking within the two theories.

but until any of these theories make successful empirical predictions their significance for interpreting QFTs must be limited.

2.4 Why physical cutoffs are also a problem for CQFT

Even though, as I have argued, there is currently no physically motivated reason for supposing cutoffs to be physical, it may be the case that we find such a reason in the future. Perhaps we will need radically different methods from those of field theory to describe physics beyond the Standard Model. There is no shortage of candidates that claim to radically alter our picture of the world—from 11-dimensional string theory to discrete spacetime to the emergent spacetime of loop quantum gravity. Though experimental support for any of these speculative theories would mean that the axioms of any AQFT must be at best only approximations, this does not mean that CQFT would escape unscathed. Any observed violation of Lorentz invariance would signal bad news for both AQFT and CQFT, and the extent to which we choose to reject or salvage the former, we should do the same for the latter.

Though its importance is not encoded in a set of axioms, Poincaré invariance is of central importance to the physical content of CQFT. In constructing QFTs, one starts by writing down a classical Lagrangian to encode the physical content of the theory. The two major constraints on the form of candidate Lagrangians are renormalizability (dealt with above) and Poincaré invariance. Since the Lagrangian is a scalar, it must remain strictly invariant under the action of the Poincaré group on its component fields. All of the fundamental forces—as described by the Standard Model—are encoded in Lagrangians obeying strict Poincaré invariance. If anything qualifies as physically relevant to CQFT, the Lagrangian certainly does; it is the starting point for building a

QFT, and determines the types of fields, their masses, and the particulars of their interactions. A violation of Poincaré invariance at a more fundamental level—be it in a particular physical process or in the structure of some new spacetime picture—undercuts to the same extent the physical significance of *any and all* theories that depend on Poincaré invariance for their formulation. Thus, despite the lack of rigid and precise axioms demanding Poincaré invariance, the physical content of CQFT stands or falls with AQFT.[10]

Once again, the major difference between AQFT and CQFT lies in the formalism. Though the *physical* content of CQFT is built upon Poincaré invariance[11], the formalism is indifferent to the constraints placed upon the Lagrangian. The success of field theoretic methods in CMP is evidence of the flexibility of the formalism; in CMP the Galilean group is taken as the appropriate symmetry group, given the low energies dealt with. In contrast, the formalisms of various AQFTs are constructed around the axioms. Any theorems that rely on exact Poincaré invariance will only hold in the real world if nature is Poincaré invariant.[12] The greater precision of the formalism in AQFT makes it more rigid in this regard.

If violations of Poincaré invariance are problematic for all variants of QFT, should investigators into the foundations of QFT fret if such violations are experimentally

[10]CQFT *methods* could still be useful, but the theoretical framework of CQFT—as encoded in the Standard Model—depends on Poincaré invariance.

[11]Depending on how one views Poincaré invariance, this may seem odd. The specific transformation properties of scalars, vectors, and tensors under the Poincaré group are undoubtedly formal properties of the particular field representations. However, the physical symmetries represented in this way have a physical basis (e.g., rotation invariance implies that the physical system can be modelled the same way when rotated).

[12]Though it isn't always possible, proofs of the form "If Minkowski spacetime then x" are strengthened and made more robust by also showing "If *approximately* Minkowski spacetime then *approximately* x. Given that our best current theories lead us to believe that spacetime is only locally Minkowski, these are the results for which we can have a high degree of confidence in their robustness.

confirmed? No; the experimental success of QFT implies that the world is at least *approximately* Poincaré invariant, and any evidence revealing the limits of that approximation has no bearing on the theory itself. We have good reason to believe that the QFTs in the Standard Model are not the final story: General Relativity implies that strong gravitational effects distort spacetime, and that our spacetime is only ever Minkowski in small patches where gravity is negligible. Though this approximation seems to hold for experiments at the LHC, if we want a theory that gets spacetime symmetries *exactly* correct, QFTs relying on Poincaré invariance will not do the trick. Rather than abandoning foundations of QFT for being approximate at best, investigation should proceed given that QFTs are highly successful within the energy domain currently testable. To this extent, we are justified in viewing the world as approximately described by QFTs, and should content ourselves with investigating an incomplete (though highly accurate) picture of nature. Whether we are dealing with a formalism that encodes Poincaré invariance into its axiomatic framework, or a formalism in which Poincaré invariance has been used indirectly to construct empirically successful theories, we should not take violations of Poincaré invariance as signalling the failure of either approach. Any robust results obtained within either formalism will still hold approximately, and should be equally subject to foundational analysis.

3 Conclusions

I have tried to show that cutoffs do not provide physical grounds for separating AQFT and CQFT as rival research programs. First, RG methods can be incorporated into AQFT without major issue, and cutoffs can be introduced as well—though explicit

cutoffs provide a more pressing conceptual revision to AQFT. Second, we needn't take AQFT to be an exact description of the world. In the same way that classical fluid dynamics is compatible with classical point mechanics, AQFT defined to arbitrary precision can be compatible with a CQFT that requires cutoffs. The apropriate lesson is that we should take AQFT to be approximately true in sufficiently low energy domains. Finally, even if cutoffs are of physical significance, they don't require a breakdown of continuum methods in general. This idea stems from pushing an analogy with CMP, which appears to be unjustified.

Though the Fraser-Wallace debate has spawned increased investigations into the foundations of QFT, it has set the boundaries of the debate in such a way as to create a false dichotomy: one is forced to choose whether to immerse oneself in the AQFT or CQFT formalisms. When we discard the false dichotomy and recognize AQFT as complementary to CQFT, we open the door to the synthesis of axiomatic methods with Lagrangian QFT. In this way the general features of QFTs can be investigated rigorously in AQFT, and we can be confident that—insofar as the axioms of AQFT capture the physical assumptions of CQFT—the results carry over to CQFT.

Though it is true that there do not yet exist AQFT models that incorporate interactions in four-dimensional spacetime, the successes of AQFT have been compatible with CQFT. Free field theories and ϕ^4_2 interaction theories constructed in AQFT give predictions in agreement with comparable CQFTs. Insofar as AQFT is a successful formalism, its results should be thought of as complementary to those of CQFT: one uses the same physical principles to construct differing formalisms.

In essence, I advocate for a position similar to Wallace's earlier view (though note that in this passage he refers only to specific results of AQFT, such as the spin-statistics

theorem):

> the foundational results which have emerged from AQFT have been of considerable importance in understanding QFT and in general they apply also to Lagrangian QFTs. This paper should be read as complementary to, rather than in competition with, these results (2006, p. 35).

The particular choice of formalism will depend on the scope of the foundational investigation. If the goal is to prove general results applicable to any relativistic QFT, then AQFT is the appropriate formalism; if the goal is to determine the consequences of specific physical interactions, then CQFT should be used.

References

Buchholz, Detlev and Rainer Verch (1995). "Scaling algebras and renormalization group in algebraic quantum field theory". In: *Reviews in Mathematical Physics* 7.8, pp. 1195–1239.

Fraser, Doreen (2009). "Quantum field theory: Underdetermination, inconsistency and idealization". In: *Philosophy of Science* 76, pp. 536–567.

— (2011). "How to take particle physics seriously: A further defence of axiomatic quantum field theory". In: *Studies in History and Philosophy of Modern Physics* 42, pp. 126–135.

— (2016). "The development of renormalization group methods for particle physics: Formal analogies between classical statistical mechanics and quantum field theory". Forthcoming in The British Journal for the Philosophy of Science.

Fraser, Doreen and Adam Koberinski (2016). "The Higgs mechanism and superconductivity: A case study of formal analogies". Forthcoming in Studies in the History and Philosophy of Modern Physics.

Haag, Rudolf and Daniel Kastler (1964). "An algebraic approach to quantum field theory". In: *Journal of Mathematical Physics* 5.7, pp. 848–861.

Halvorson, Hans and Michael Müger (2007). "Algebraic quantum field theory". In: *Handbook of the Philosophy of Physics, Part A*. Ed. by Jeremy Butterfield and John Earman. Elsevier.

Ruetsche, Laura (2011). *Interpreting quantum theories*. Oxford University Press.

Wallace, David (2006). "In defence of naiveté: The conceptual status of Lagrangian quantum field theory". In: *Synthese* 151, pp. 33–80.

Wallace, David (2011). "Taking particle physics seriously: A critique of the algebraic approach to quantum field theory". In: *Studies in History and Philosophy of Modern Physics* 42, pp. 116–125.

Wilson, Kenneth (1983). "The renormalization group and critical phenomena". In: *Reviews of Modern Physics* 55.3, pp. 583–600.

Wilson, Kenneth and John Kogut (1974). "The renormalization group and the ϵ expansion". In: *Physics Reports* 12.2, pp. 77–199.

On Epistemically Detrimental Dissent:

Contingent Enabling Factors v. Stable Difference-Makers.

Soazig Le Bihan and Iheanyi Amadi

Abstract.

The aim of this paper is to critically build on Justin Biddle and Anna Leuschner's characterization (2015) of epistemologically detrimental dissent (EDD) in the context of science. We argue that the presence of non-epistemic agendas and severe non-epistemic consequences are neither necessary nor sufficient conditions for EDD to obtain. We clarify their role by arguing that they are contingent enabling factors, not stable difference-makers, in the production of EDD. We maintain that two stable difference-makers are core to the production of EDD: production of skewed science and effective public dissemination.

Introduction.

The aim of this paper is to critically build on Justin Biddle and Anna Leuschner's characterization of epistemologically detrimental dissent (EDD) in the context of science (2015). We follow their lead in taking 'dissent' to be a particular kind of criticism, i.e. the act of objecting to a widely held conclusion. When done properly, dissent is welcome within scientific practice. As Helen Longino has clearly established, "scientific knowledge is produced collectively through the clashing and

meshing of a variety of points of view (1990, 69). Criticism, when done properly, is integral to the collective advancement of science.[1] Dissent, when an instance of proper criticism, is thus epistemically valuable in the context of science.

Now there are some instances of dissent that come out as epistemically detrimental. That is to say, some instances of dissent seem to impede, not promote, the collective advancement of science. Many examples come to mind, that have been well described in the recent literature (Oreskes and Conway 2010, Biddle and Leushner 2015, Harker 2015). Roughly speaking, EDD is about manufacturing controversy in a particular scientific field. The typical story goes something like the following. The research involved has some severe non-epistemic consequences in terms of, on one side, industry profit, and, on the other side, public welfare; large amounts of money are invested by industry-related groups to (1) produce some skewed research, (2) largely publicize the results through the media, (3) produce an atmosphere of confusion and doubt within the public, (4) launch some campaign against the lead scientists of the field in the media and political world (often through personal attacks and threats); this results in an atmosphere in which the scientists subjectively feel a lot of pressure and discomfort, and also objectively waste precious time and limited resources to address the well-publicized skewed research. At this point, the collective advancement of science is clearly impeded. We have an instance of EDD.

[1] Longino (1990) offers an account of some of the various kinds of epistemically beneficial criticism within science.

The aim of this paper is to properly distinguish, in that story, between (1) contingent enabling factors, and (2) stable difference-makers, in the production of EDD. Our most contentious claim is that the intrusion of non-epistemic agendas and presence of severe non-epistemic risks are contingent enabling factors, not stable difference-makers for EDD. We maintain that two stable difference-makers are core to the production of EDD: production of skewed science and effective public dissemination.

In Section 1, we offer what we take to be the most straightforward argument for the claim that intrusion of non-epistemic agendas is not sufficient in the production of EDD: it may lead to EDD only if it leads to skewed science. In Section 2 we argue that it is not necessary either. Section 3 is devoted to a clarification of the role of intrusion of non-epistemic agendas in EDD on the basis of a distinction between contingent enabling factors and stable difference-makers. Section 4 investigates the consequences of our analysis for the Inductive Risk Account of EDD proposed by Biddle and Leuschner (2015).

Section 1. Non-epistemic agendas: not sufficient for EDD

That intrusion of non-epistemic agendas is not sufficient to the production of EDD has been discussed by Wilholt (2009), and Biddle and Leuschner (2015). Roughly, the point is simply that, unless intrusion of background non-epistemic agendas is such that the work produced *fails to satisfy some of the conventional standards for proper science*, there is no problem. We offer here what we take to be the most straightforward argument for this point.

As the community of philosophers of science have recently come to recognize, intrusion of non-epistemic values in scientific practice is quite common (Douglas 2009). Now obviously, that does not necessarily result in skewed science. If a scientist defends a conclusion C on the basis of evidence E, the fact that some background non-epistemic values enters in her reasoning does not matter if (1) she can publicly produce a reasoning in defense of C, and if (2) that reasoning can be assessed as adequate scientific reasoning by her peers, including peers who do not share the same background non-epistemic values. If these two conditions are met, then the conventional standards for proper science are met, and we do not have a case of skewed science. Now if proper scientific work was produced, there is no a priori reason to think that her work cannot partake in the collective advancement of scientific knowledge. It might do so at various degrees, but that will depend on its heuristic value, which is a priori unrelated to whether or not there was intrusion of non-epistemic values.

Let us push this line of argument a little further. It is important here to underline the fact that the reasoning rendered public by the scientist might not be the actual reasoning through which she came to accept either E or its relevance with regard to C. From a subjective point of view, for example, she might well have had accepted C well before she produced E and the reasoning defending the relevance of E as supporting C. She might well have accepted C for non-epistemic, value-laden, reasons. However, such considerations over the subjective state of scientists do not matter. The collective assessment of scientific research is not in the business of mind reading. No matter what kind of reasoning (or non-reasoning) actually

brought a scientist to believe C, the relevant question is whether she is capable of producing a reasoning in defense of E and its relevance with regard to C that can be publicly, and positively, assessed by the experts in her field. To put it bluntly: the most biased and ill-intentioned scientists are a priori capable of producing good scientific work.[2]

This line of argument applies to the production of dissenting views. Dissenting claims proposed by scientists motivated by non-epistemic agendas do not necessarily lead to skewed science and hence to of EDD. If a reasoning can be publicly produced, and if the members of the scientific community, including members of that community who do not share the same values as the dissenting views' proponents, assess that reasoning as scientifically adequate, then we do not have an instance of skewed dissent. As an instance of work that satisfies the agreed-upon standards of proper scientific practice, the dissenting view could well participate in the advancement of scientific knowledge. It could do so at various degrees, depending on how important the dissenting views are, but that would not depend on whether or not the dissenting views are the product of scientists with non-epistemic agendas. Considerations about the subjective intentions, or background beliefs, of the scientists are irrelevant, unless one can show that skewed science was produced.

[2] This is not denying the actuality of implicit bias. By definition, implicit bias is still bias. As such, it can be recognized by the scientific community for what it is. What is implicit about it is that the biased author (and possibly some of her peers as well) is not even realizing her own bias.

Section 2 Non-epistemic agendas: not necessary for EDD

At this point, we have shown that intrusion of non-epistemic agendas do not necessarily result in the production of EDD. Note that EDD does not require intrusion of non-epistemic agendas either. What would it take to have a case of EDD without any intrusion of non-epistemic agendas? We know that EDD is about manufacturing controversy within a scientific field. First, the controversy is "manufactured", not genuine, because the dissenting view is not based on proper science; it violates some of the commonly accepted standards for proper scientific practice; it is an instance of skewed science. Now skewed science can come to be in many ways. It does not have to result from the intrusion of non-epistemic agendas. One can imagine the case of a scientist, say Jack, who is genuinely interested in partaking in the collective advancement of scientific knowledge, but is also a poor scientist. One can imagine that Jack is very wealthy, and thus has both the time and financial resources to pursue his research, and produce a large amount of work challenging the commonly held views in a given scientific field. Jack, albeit misguided in many ways, could conceivably do all of this with the "purest" goal in mind.

Now one immediately sees that the production of bad science is not enough to produce EDD. Jack's research is likely to be simply ignored by the scientific community. So what would it take to "manufacture" a controversy on the basis of Jack's research? The answer seems rather straightforward: Jack's research needs to be effectively disseminated, so that scientists feel pressured to respond to Jack's

challenges. The standard avenues for dissemination of scientific research, i.e. peer-reviewed publication, however, are not likely to be an option for Jack, since his work is widely recognized by the community as being of poor scientific quality. He must then bypass these avenues, and manage to effectively disseminate his research among the public. Mass media would be a likely option for this. This in turn forces scientists in the field to waste time and resources to address Jack's research. Hence a case of EDD, with the purest epistemic goal at its source.

The case above might seem far-fetched. One objection could be that, unless some non-epistemic values were at stake, it is unlikely that the media and the public would get interested in Jack's research, and Jack would fail to be able to manufacture the controversy. It might be unlikely, but it is surely conceivable. If Jack's public dissemination machinery is effective enough, (mis-) understandings over the state of research in the field of concern could well have serious repercussions on public funding. Jack could well have a very strong network of communication – he could well be the owner of a very large cable and press network. Repeated reporting on public funding of supposedly controversial science could well spur outrage in the public. "Debates" on mass media would ensue. As soon as the scientists would engage in that conversation, Jack's claims would gain in credibility.[3] At the end, Jack's campaign could well be so effective that scientists

[3] This is a point that Hannah Arendt made clear in her insightful analysis of controversy- and doubt-manufacturing in a completely different context, i.e. the (non-)issue of the reality of the Holocaust during WWII (1966/2010).

would indeed be forced to repeatedly address his research to defend their own. So, intrusion of non-epistemic agendas is not necessary to the production of EDD.

Section 3. Stable Difference-Makers v. Contingent Enabling Factor

From the discussion above, we conclude that intrusion of non-epistemic agendas is neither necessary nor sufficient for the production of EDD. Such a conclusion might strike many as unsatisfactory, however. Isn't it the case that intrusion of non-epistemic agendas was an important factor in the production of the common cases of EDD that we have witnessed over the last 50 years? Some may even want to claim that, as a matter of fact, in all of the cases we know of in recent history, no EDD would have occurred if it were not for the intrusion of non-epistemic agendas. This is an important intuition, and arguably, any satisfactory account of EDD ought to make sense of it. Fortunately, we believe there is a way to do so, that is, by appealing to the distinction between contingent enabling factors and stable difference-makers as discussed by Thomson (2003) and Woodward (2010). Thomson (2003) makes the point (contra many theories of causation) that just because 'E would not have happened without C', it does not follow that 'C has caused E'. She argues that the proposition 'E would not have happened without C' only entails that 'C was physically necessary for E'. Consider her example. John built a bridge over the Rapid River. The Rapid River is notoriously wild, and only John, a master-builder, could have done it. From the bridge being built, it ensues that Smith crosses the river. Now John's building the bridge was physically necessary to Smith's crossing the Rapid River, but most would agree that it is misguided to take it

as a cause for it. John's building the bridge, even if "physically necessary" in the whole process, remains largely irrelevant to Smith's crossing the river. It belongs to the background conditions, or environmental conditions, that make Smith's crossing possible, without causing it in any genuine sense of causation. In Thomson's vocabulary, it is only an enabling factor.

Woodward (2010) is interested in analyzing a similar distinction between the core difference-makers and the background conditions. His analysis is useful to flesh out some of the characteristics of enabling factors à la Thomson.[4] One of intuitions Woodward is trying to capture is that some causal relationships are robust, i.e. insensitive to environmental change, while others are contingent on the presence of a specific environment. To do so, he articulates the notions of "stability".[5] A causal relationship, according to Woodward, is stable if and only if it holds over a wide range of background conditions. Some examples might be useful at this point.

[4] Note that we do not claim (and neither does Woodward) to have unveiled the set of necessary and sufficient conditions for factors to qualify as enabling factors by contrast to stable difference-makers. We will only claim that being enabling factors are typically unstable, and hence, that lack of stability serves as a good indicator for a factor to be only enabling, not causing.

[5] Two other notions are articulated in the article. The notion of proportionality serves to address the issue of the proper levels of explanation. The notion of specificity serves to address the issue of coarse v. fine-grain causal influence.

A paradigmatic example of an unstable relation would be the following.[6] "Star" professor P writes a letter of recommendation for Jane, thanks to which Jane gets a job at university U. She would not have gotten the job without it. Jane meets Joe at U, they get married, and have children. Challenged by the difficulties of coupling an academic career with quality parenting, Jane goes into depression. Now consider the following claim: 'P's writing a letter for Jane caused Jane's depression'. Given the story that is given, there is a sense in which P's writing a letter for Jane enabled Jane's suffering from depression, but there is also a strong sense in which it is misguided to take it as a cause for it. The reason is that the relation between P's writing the letter and Jane's suffering from the disease would cease to hold under many small, contingent, changes in the background conditions for the story (Jane and Joe could not have met, they could have decided to not have children, U could have had a very progressive parental leave policy, etc.). The causal relationship between the letter and the depression is thus highly unstable because it holds only in a very specific environment.

Now contrast this with a paradigmatic example of a stable relation. I turn on the heat under my closed pressure cooker (with some water in it). The pressure goes up and the valve shuts down. Clearly, heating up the pressure cooker is a stable cause of the pressure valve to shut down. Many of the most stable causal relations are backed up by what the kind of generalizations that we take to be the laws of physics, or chemistry. These generalizations hold over a wide range of background conditions.

[6] This example is inspired by Woodward (2010) himself inspired by Lewis (1986).

There are obviously various degrees of stability in between these two extreme cases. Stability is not an all or nothing affair. It might also be difficult to figure out which causal relationships are more or less stable. That said, it could also be worth the effort looking into it, because, how stable a factor is could be a measure of how well we can target change by targeting that factor in a given situation. As Woodward explains (2010, 315): "other things being equal, causal relationships that are more stable are likely to be more useful for many purposes associated with manipulation and control than less stable relationships." Applied to our case, if ultimately we hope to be able to alter the manufacturing of controversy and EDD, it could turn out to be very useful to clarify the causal landscape behind EDD by distinguishing between the contingent enabling factors and the more stable difference-makers.

Thomson's and Woodward's analyses are clearly related. Thomson's bridge example is a clear case of a very unstable causal relationship: it holds only under very specific background conditions (The Rapid River could have been gently, Smith could have decided not to cross the bridge, etc.) Some unstable causal relationships as discussed by Woodward are so at least partially because they are relationships of contingent "physical necessity" à la Thomson. So, a causal factor may be highly unstable, despite being 'necessary' to the causal process, if its influence on the process is highly contingent on a specific environment. No matter how "necessary"

in that sense a factor F is, F being unstable points F being a enabling factor, not a stable difference-maker.[7]

The discussion above allows us to bring home two important points. First, it allows us to identify two stable difference-makers for the production of EDD: the production of skewed scientific research and its effective public dissemination. That the combination of these two factors produces an instance EDD holds over a wide range of conditions. What changes in background conditions would make that causal relation to fail? First, one could think of a world in which scientists could ignore even well-advertised skewed science. For example, that could possibly be the case in a world in which production of scientific research would not depend on getting public founding, or in a world in which the public is generally knowledgeable about (the philosophy of) science, and hence, is able to recognize that the well-

[7] Two points of clarification are in order. First, Woodward convincingly argues that the extent to which a cause is stable is related, but not equivalent to, its distal/proximate character vis à vis the effect. Second, Woodward also argues that stability is not dependent on the level of explanation: degrees of stability are not necessarily to how "reductive" the explanation is. So, our distinction between contingent enabling factors and stable difference-makers is not trivial in the sense that the most stable difference-makers would always be the most proximate causes described at the level of fundamental particles.

advertised science is skewed. Arguably, these do not qualify as small changes in the background conditions for scientific practice.[8]

The second point is a clarification of the role played by the intrusion of non-epistemic agendas in the production of EDD. Intrusion of non-epistemic agendas is not a stable difference-maker for the production of EDD. This is because there is a large range of conditions under which intrusion of non-epistemic agendas do not result in EDD. These include the conditions for all the cases in which intrusion of non-epistemic agendas do no result in skewed science. If we take seriously recent work on science and value, intrusion of non-epistemic values is actually the rule, not the exception within the practice of science (Douglas 2009, Intemann 2001, 2015, and references therein). Note that, if our take on Thomson's and Woodward's analyses is correct, then the claim that intrusion of non-epistemic agendas is not a stable difference-maker but only a contingent enabling factor is consistent with the fact that it has been "physically necessary" in many of the well-known instances of EDD. One can consistently say that, while not a stable difference-maker, it has been an important enabling factor for the production of well-publicized skewed science. Intrusion of non-epistemic agendas has been necessary for some groups to develop an *interest in funding* the production and public dissemination of skewed research.

[8] There is also a possibility that some cases of EDD could come out of seemingly proper science "distracting" the public from the most widely held views within the scientific community. We believe that even in these cases, dissenting views do not entail EDD unless there is violation of some conventional standards for proper science. This interesting issue belongs to another paper.

That said it is important to distinguish between factors that are characterized by this kind of 'necessity' (the bridge or letter kind of necessity) and factors that are true stable difference-makers. It is all the more important that, if one of our goals is to alter the production of EDD, then our analysis suggests that intrusion of non-epistemic agendas is not the proper target. Once again, non-epistemic values are the common rule within the practice of science. A more efficient approach in the prevention of EDD would be to understand the various ways skewed science may be produced. This includes the important discussion on the distinction between legitimate and illegitimate use of non-epistemic values in scientific practice (Hicks 2014, Intemann 2015). This in turn includes an investigation of the mechanisms by which intrusion of non-epistemic values does result in skewed science. Implicit bias might one of these mechanisms. Inductive risk bias, as we shall explain in the next section, is another one. Before we turn to this point, let us take stock.

We have clarified the causal landscape for the production of EDD. We have identified two stable difference-makers – production of skewed science and its effective public dissemination; and we have characterized the important role of intrusion of non-epistemic agendas within science as contingent enabling factors for the production and dissemination of skewed research, hence for EDD.

Section 4. Consequences for the Inductive Risk Account of EDD

Biddle and Leuschner have articulated what they call the "inductive risk account" of EDD (2015). According to this account, the following set of conditions are jointly sufficient for the production of EDD (2015, 273):

Dissent from a hypothesis H is epistemically detrimental if each of the following obtains:

(1) The non-epistemic consequences of wrongly rejecting H are likely to be severe

(2) The dissenting research that constitutes the objection violates established conventional standards.

(3) The dissenting research involves intolerance for producer risks at the expense of public risks.

(4) Producer risks and public risks fall largely upon different parties.

Biddle and Leushner admit that these conditions are not necessarily related to the production of EDD (275):

"We are not arguing that, in all possible worlds, research that meets the conditions of the inductive risk account inhibits the progress of science. It is possible, for example, to organize science and to regulate industry in such a way that dissent that meets these conditions is not widely disseminated, does not acquire political authority, and is not used to attack mainstream scientists. But this is not the way in which science and society are currently organized. Dissent that meets the conditions of the inductive risk account is, given current societal arrangements, likely to inhibit knowledge production, particularly because of the success of political, economic, and ideological interests in structuring the dissemination of research."

We think that the framework used in Section 3 can help clarify the causal landscape for the production of EDD offered in the Inductive Risk Account. Our contention is that Biddle and Leuschner, by focusing on inductive risk, have identified a

particular, important, but still contingent, enabling factor, but have failed to clearly distinguish the proper core of stable difference-makers, for the production of EDD. Let us make that point in more details.

The four conditions above can be seen as dividing into three groups. Condition (2) identifies one of the stable difference-makers – production of skewed science. Conditions (1) and (4) together specify some particular enabling conditions for the formation of non-epistemic agendas – the presence of severe and opposing non-epistemic consequences (SONEC). Condition (3) identifies a mechanism by which intrusion of SONEC-related non-epistemic agendas may enable the production of skewed science. In other words, the inductive risk account of EDD identifies an important series of enabling causes leading to one of the two stable difference-makers we have identified in Section 1-3, i.e. production of skewed science. That series of cause is something like this: from the presence of SONEC to biased inductive risk reasoning, and to skewed science. This is an important contribution to the understanding of EDD precisely because it not only identifies some particular enabling factors (the presence of SONEC) for the formation of epistemic agendas, but also a mechanism by which intrusion of SONEC-related non-epistemic agendas may enable the production of skewed science (via inductive risk bias). Now it is also important to clarify the causal landscape and recognize that fulfillment of Condition (2) is the stable difference-maker which fulfillment of Conditions (1), (4), and then (3) enable as a matter of contingent fact. Biddle and Leuschner seem to have missed that useful distinction.

If our analysis in Section 3 is correct, they also have failed to include the second stable difference-maker for EDD, i.e. effective public dissemination. As they admit in the paper (see quote above), the presence of SONEC obviously does not imply that effective public dissemination will ensue. Conversely, as Jack's case shows, effective public dissemination could well be obtained without the presence of SONEC. How (un-)likely this is obviously is an empirical question. No matter how unlikely, however, it is important for our understanding of EDD to mention effective public dissemination as a core stable difference-maker. The inductive risk account fails to do so. Let us underscore, however, that Biddle and Leuschner once again have identified an important mechanism by which presence of SONEC enables effective public dissemination and the manufacturing of controversy: the presence of SONEC not only enables the production of skewed science, but also the establishment of "sophisticated, private-funded network for disseminating [dissenting] results" (2015, 275).

This brings us to our conclusion on the Inductive Risk Account: Biddle and Leuschner have successfully identified an important contingent enabling factor for EDD, i.e. the presence and influence of SONEC. That said, they have failed to distinguish between the different roles that enabling factors and stable difference-makers play in the production of EDD. We hope to have clarified the situation.

Conclusion

Well-known cases of EDD seem to have in common various forms of intrusion of non-epistemic, often SONEC-related, agendas within the science. We have argued

that such intrusion is not core to the production of EDD: neither necessary nor sufficient, it is also not a stable difference-maker. We have clarified its causal role: intrusion of non-epistemic agendas is a contingent enabling factor. Reduced to its core, EDD is just well-advertised bad science. Because it is well advertised, it has an impact on the collective building of scientific knowledge. Because it is bad science, it does not advance that endeavor, but any case negatively impacts it instead.

To make the distinction between contingent enabling factors and stable difference-makers is important for at least three reasons. First, it is important to clarify the causal landscape that leads to the production of EDD, as it simply increases our understanding of EDD. Second, it might suggest more efficient avenues for targeting change. Finally, it is crucial to make room for the intrusion of non-epistemic values within the science without it being epistemologically detrimental. As the community of philosophers of science comes to recognize that such intrusion is the rule rather than the exception, one must leave conceptual room for a distinction between "legitimate" and "illegitimate" role for non-epistemic values within science (Hick 2014, Intemann 2015).

Bibliography

Arendt, Hannah. 1967/2010. "Truth and Politics." In José Medina and David Wood (eds). *Truth. Engagements Across Philosophical Traditions*. Blackwell: 295-314.

Biddle, Justin B. and Anna Leuschner. 2015. "Climate Skepticism and the Manufacture of Doubt: Can Dissent in Science be Epistemically Detrimental?" *European Journal for Philosophy of Science* 5 (3): 261-278.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

Harker, David. 2015. *Creating Scientific Controversies: Uncertainty and Bias in Science and Society*. Cambridge University Press.

Hicks, Daniel J. 2014. "A New Direction for Science and Values." *Synthese* 191 (14): 3271-3295.

Intemann, Kristen. 2015. "Distinguishing between Legitimate and Illegitimate Values in Climate Modeling." *European Journal for Philosophy of Science* 5 (2): 217-232.

———. 2001. "Science and Values: Are Value Judgments always Irrelevant to the Justification of Scientific Claims?" *Philosophy of Science*: S518.

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Lewis, David. 1986. "Postscript c to 'causation': (insensitive causation)" in: *Philosophical papers*, vol 2. Oxford University Press, Oxford: 184–188

Oreskes, Naomi and Erik M. Conway. 2010. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Publishing USA.

Thomson, Judith Jarvis. 2003. "Causation: Omissions." *Philosophy and Phenomenological Research* 66 (1): 81-103.

Wilholt, Torsten. 2009. "Bias and Values in Scientific Research." *Studies in History and Philosophy of Science* Part A 40 (1): 92-101.

Woodward, James. 2010. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biology & Philosophy* 25 (3): 287-318.

Literal vs. careful interpretations of scientific theories: the vacuum approach to the problem of motion in general relativity

Dennis Lehmkuhl
Einstein Papers Project and HSS Division,
California Institute of Technology
Email: lehmkuhl@caltech.edu

Forthcoming in *Philosophy of Science* (PSA 2016 Supplement)
Version: September 26, 2016

Abstract

The problem of motion in general relativity is about how exactly the gravitational field equations, the Einstein equations, are related to the equations of motion of material bodies subject to gravitational fields. This paper compares two approaches to derive the geodesic motion of (test) matter from the field equations: 'the T approach' and 'the vacuum approach'. The latter approach has been dismissed by philosophers of physics because it apparently represents material bodies by singularities. I shall argue that a careful interpretation of the approach shows that it does not depend on introducing singularities at all, and that it holds at least as much promise as the T approach. I conclude with some general lessons about careful vs. literal interpretations of scientific theories.

Contents

1 Introduction

It is a bit of an irony that one of the most widely embraced definitions of what it means to be a scientific realist is due to the arch-anti-realist Bas van Fraassen. His definition starts by stating that "Science aims to give us, in its theories, a literally true story of what the world is like".[1] And indeed, scientific realists often see themselves as committed to 'taking scientific theories at face value': if the best theories of particle physics say that quarks exist, then we should believe that they exist; if general relativity tells us that gravity is really just an aspect of spacetime structure, then we should believe it; if quantum mechanics tells us that the world is at its core non-deterministic, then we should believe that too.

The problem is that scientific theories, or at least the theories of modern physics, are not that straightforward with us. They may seem so at first, but if you listen to the details of their respective stories, if you take your time to look under the surface, what exactly we should take them to tell us about the world is far from clear. Murray Gell-Mann, the inventor of the concept of quarks, for a long time did not think that quarks should be interpreted as literally existing; neither did Richard Feynman. Albert Einstein passionately resisted the interpretation of general relativity that says that the gravitational force field of Newtonian theory is ontologically reduced to the geometry of spacetime in general relativity. And of course, there is a long-standing battle in foundations of physics about whether quantum mechanics really does tell us that the world is non-deterministic.[2]

In this paper I shall introduce a new case study that provides further evidence for the position that, whether you are a realist or not, the *literal interpretation* of a scientific theory, especially in physics, can be rather misleading. I will argue that what we should aim for is a *careful interpretation*;

[1] Van Fraassen [1980], p.8.

[2] For a discussion of different interpretations of the quark concept see Pickering [1999], for Einstein's opposition to interpreting general relativity as a geometrization of gravity see Lehmkuhl [2014], and for debate on whether quantum mechanics is really indeterministic see e.g. Saunders et al. [2010].

an interpretation of the theory or model or formalism that engages with its details, both with the details of its mathematical structure and with how it is applied to the natural world. Philosophy of science must be willing to look under the hood.

The case study I want to look at is the so-called problem of motion in the general theory of relativity (GR). It asks about the precise relationship between the two sets of equations that are at the very heart of GR. On the one hand there are the Einstein field equations, which give us the dynamics of the gravitational potential (the metric tensor) $g_{\mu\nu}$:

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} = \kappa_E T_{\mu\nu} \quad . \tag{1}$$

On the other hand, we have the geodesic equation that determines which paths through spacetime are geodesics of the connection $\Gamma^{\nu}{}_{\mu\sigma}$ compatible with the metric $g_{\mu\nu}$:

$$\frac{d^2 x_\tau}{ds^2} + \Gamma^{\tau}{}_{\mu\nu} \frac{dx_\mu}{ds} \frac{dx_\nu}{ds} = 0. \tag{2}$$

In GR, material bodies subject only to gravitational fields are supposed to move on the geodesics determined by equation (2).[3] The problem of motion in GR is the question of whether the equations of motion of matter subject to gravitational fields (2) can be derived from the gravitational field equations (1).

Einstein himself, in his first publication on the topic, a paper co-written with Jakob Grommer and published in 1927, compares different classes of attempts to give such a derivation. In particular, Einstein and Grommer distinguish between two classes of attempts at deriving the geodesic motion of matter from the gravitational field equations, which I will term *the T approach* and *the vacuum approach*, respectively. The T approach starts from the realization that the field equations (1) imply the conservation condition, namely that the covariant divergence of the energy-momentum tensor $T_{\mu\nu}$ vanishes:

$$\nabla^{\mu} T_{\mu\nu} = 0 \quad . \tag{3}$$

[3]It is a big question which systems are actually included under 'material bodies' here. The minimal position is that only test particles are referred to: particles with negligible extension, spin, and self-gravity. However, many actual bodies can be approximated well by test particles in this sense; planets orbiting a star are an example, as we shall see below.

From this, together with certain conditions on the energy-momentum tensor $T_{\mu\nu}$, the T approach derives that material particles move on time-like geodesics. It is this kind of approach to the problem of motion that philosophers have engaged with almost exclusively up to now.[4]

Einstein and Grommer end up dismissing the T approach, and suggest an alternative path to deriving geodesic motion instead. It is a particular version of a *vacuum approach to the problem of motion*. Einstein and Grommer start from the vacuum form of the Einstein field equations,

$$R_{\mu\nu} = 0 \quad , \tag{4}$$

and attempt to derive that the equations (4) imply that material particles move on geodesics.

To the extent that philosophers have engaged with this approach at all, they have quickly dismissed it because it seems to model material bodies by singularities in spacetime; while singularities, by definition, are not even part of spacetime. However, in this paper I shall argue that this dismissal was far too fast, and that indeed the vacuum approach deserves at least as much attention by philosophers as the T approach. The vacuum approach, despite first appearances, engages more closely with some of the most major predictions of GR: both the prediction of the perihelion of Mercury and the prediction of light bending by the Sun utilise the vacuum approach to the derivation of motion of material systems. Indeed, even the prediction of gravitational waves resulting from a binary black hole merger that was recently confirmed rests on the vacuum field equations, for black holes are described by vacuum solutions.[5]

My argument in this paper will proceed in three steps. First, I will argue that the vacuum approach to the problem of motion promises certain advantages that the T approach lacks. Second, I will argue that the problems of the vacuum approach for which it has been dismissed are artefacts of a too literal interpretation of the formalism and its application to the problem at hand. Third, I will argue that a careful interpretation makes the problems disappear; I will argue that the approach does not need to interpret singularities as representing material bodies.

[4]For a comprehensive review of the early history of this approach see Havas [1989] and Kennefick [2005]; for two particularly beautiful exemplars from within this class of proofs see Geroch and Jang [1975] and Ehlers and Geroch [2004], which are investigated by Brown [2007], Malament [2012], and Weatherall [Forthcoming, 2011].

[5]See Abbott et al. [2016] and references therein.

2 A critical comparison of the two research programmes

I said above that the T approach to the problem of motion proceeds via the fact that the Einstein field equations (1) imply the conservation condition (3), which in turn implies the geodesic motion of matter. However, as Malament [2012] pointed out, the conservation condition by itself is not sufficient to prove that the geodesic equation is the equation of motion of material particles. One of the most general proofs from within the T approach, proposed by Geroch and Jang [1975] and further generalised by Ehlers and Geroch [2004], rests not only on the conservation condition (3), but also on the strengthened dominant energy condition, which states:

> Given any timelike covector ξ_μ at any point in M, $T^{\mu\nu}\xi_\mu\xi_\nu \geq 0$ and either $T^{\mu\nu} = \mathbf{0}$ or $T^{\mu\nu}\xi_\mu$ is timelike.

The first clause is effectively the weak energy condition, which states that the mass-energy-momentum density associated with the body in question is always non-negative. The second clause states that every observer will judge the mass-energy-momentum of the body to propagate along time-like curves only.[6]

It would be rather attractive if we did not have to presume that material particles move on time-like curves to then show that these curves are actually time-like geodesics, and if we did not have to presume that matter cannot have non-negative mass-energy. These are weak assumptions about the nature of matter, but they are assumptions.

The vacuum approach to the problem of motion, on the other hand, aims to make *no* assumptions about the nature of matter and its properties at all, and to still derive that matter moves on geodesics. It starts from the question of whether just knowing the exterior gravitational field of a material body, and how this gravitational field interacts with the gravitational field of its surroundings, is enough to derive that the body will move on a geodesic of the metric surrounding it. Arguably, this programme is far more ambitious than the T approach, for it starts with fewer assumptions.[7] And yet, if successful, it would really fit much better the virtues that philosophers have associated

[6]For more on the interpretation of the strengthened dominant energy condition seeWeatherall [2011], Weatherall [Forthcoming] and especially Curiel [Forthcoming].

[7]One might be tempted to argue that despite first appearances the vacuum approach

with the geodesic theorem(s) in the first place: deriving the inertial motion of matter from knowledge of the dynamics of gravitational fields alone.[8]

Einstein was deeply skeptical of the role of the energy-momentum tensor in GR. Throughout the decades, he emphasised that $T_{\mu\nu}$ provides only a 'phenomenological representation of matter'.[9] In Einstein and Grommer [1927], Einstein elaborates that general relativity with an energy-momentum tensor as a source term on the right-hand side of (1) is just not a complete theory: it does not tell us what kind of matter is present, only that it has a certain mass-energy distribution. This perspective on GR was further strengthened by Tupper [1981, 1982, 1983], who showed that knowing the energy-momentum tensor of a material system does not suffice to tell us what kind of matter is present. For example, one and the same mass-energy-momentum distribution $T_{\mu\nu}$ featuring on the right-hand side of the Einstein equations, and solving the Einstein equations for the same metric, can correspond either to an electromagnetic field or a viscous fluid. Knowing the energy-momentum tensor is just not sufficient to know which of these two material systems it is that interacts with the metric field.

Einstein's aim is then to instead start with the vacuum field equations

starts with more demanding assumptions than the T approach. For the vacuum Einstein equations (4) logically imply that the strengthened dominant energy condition (SDEC) holds for the Ricci tensor $R_{\mu\nu}$. The opposite is not true, so that demanding Ricci flatness is clearly a stronger constraint on the Ricci tensor than demanding that it obeys the SDEC. But concluding from this that the vacuum approach starts from stronger assumptions than the T approach would be a mistake. For the T approach assumes i.) the full Einstein field equations (1); and ii.) that the energy-momentum tensor (and thus the Einstein tensor) adheres to the SDEC. The vacuum approach only assumes the vacuum Einstein equations (4), and thus starts with weaker assumptions than the T approach. However, it might well be that depsite *starting* with weaker assumptions than the T approach, a particular manifestation of the vacuum approach might end up with stronger assumptions than a particular manifestation of the T approach. For example, the 1927 Einstein-Grommer vacuum approach, discussed below, involves, among other demands, a so-called equilibrium condition which is supposed to relate solutions to the non-linear field equations to solutions of the linearized field equations in a particular way; no such demand is included in, say, the Geroch-Jang version of the T approach. Thus, further analysis might well show that Einstein and Grommer use stronger assumptions than Geroch and Jang. Einstein himself would likely have been content with that, as long as it allowed him to avoid the introduction of $T_{\mu\nu}$, for reasons discussed below.

[8]Cf. Brown [2007], p. 141 and 163.

[9]See, for example, Einstein [1922], Einstein to Michele Besso, 11 August 1926 (EA-7-361), and Einstein [1936].

(4), treat material particles as singularities in the metric field,[10] and derive that they move on geodesics of a metric $g_{\mu\nu}$ that solves the vacuum field equations (4) in the region through which the particle moves.

To the extent that philosophers have engaged with this approach at all, they have already dismissed it at this point. The main criticism is that the very idea of the approach is flawed: A singularity is not even part of spacetime. How should it be possible to describe its motion in said spacetime?

Both Torretti and Earman essentially answer that this is not possible and that the whole programme is ill-conceived. Earman [1995], p. 12, writes:[11]

> [S]ingularities in the spacetime metric cannot be regarded as taking place at points of the spacetime manifold M. Thus, to speak of singularities in $g_{\mu\nu}$ as geodesics of the spacetime is to speak in oxymorons.

The most detailed discussion of the Einstein-Grommer paper in the philosophical literature is due to Tamir [2012]. After quoting the above statement by Earman, Tamir goes on to write (p.142):

> The proponent of such a "vacuum-cum-singularity" technique is faced with the rather paradoxical challenge of explaining in what sense we can say that a singular curve (ostensibly constituted by the *missing* points in the manifold) is actually a geodesic of the spacetime from which it is absent. Not only is no metric defined at the singularity, but also technically there are not even spacetime points there: the geodesic does not exist.

Tamir then mentions a key ingredient of the Einstein-Grommer approach, namely the distinction between an 'inner metric' and an 'outer metric'.[12] Einstein and Grommer aim to show that the particle characterized by a

[10] In recent years, the adequate definition of a singularity in GR has been a subject of extensive debate, see e.g. Earman [1995] and Curiel [1999]. For Einstein's thoughts on singularities see Earman and Eisenstaedt [1999]; in the context of the Einstein-Grommer paper Einstein clearly thinks of a singularity in the metric field $g_{\mu\nu}$ as a region where the components of the metric tend to infinity.

[11] For similar statements see Torretti [1996], section 5.8.

[12] There is an interesting relationship between Einstein and Grommer's distinction between inner and outer metric (discussed further in section 3) on the one hand and the later distinction between interior and exterior black hole solutions on the other. I do believe that bringing together results and concepts developed in the context of black hole solu-

singular inner metric moves on geodesics of the non-singular outer metric. Tamir states that the "suggested implication" is that we are to compare a second spacetime whose metric is that of the regular outer metric with the singular first spacetime, and identify the regular geodesic of the second spacetime with the singular curve of the first one. He then argues that the thought that the second singularity-free spacetime can teach us anything about the singular original spacetime is "spurious".

My point in the following will be this. Even if this argument were convincing, its premise (the 'suggested implication' that Einstein and Grommer intended to deduce something about a singular spacetime by comparing it to a non-singular spacetime) is not. I shall argue that by looking at the details of the Einstein-Grommer approach we come to a different interpretation of the approach, one that sheds a completely different light on the alleged presence of singularities. We will see that a careful (rather than literal) interpretation of the vacuum approach, and the Einstein-Grommer paper in particular, does not actually depend on introducing singularities at all.

3 The vacuum approach to the problem of motion

3.1 Two ways of looking at Einstein's model of the Sun-Mercury system

In a way, the story of the vacuum approach to the problem of motion starts in 1915, with Einstein's treatment of the orbit of Mercury around the Sun in the context of GR. It is a two-body problem: a small body (Mercury) with a comparatively small mass orbits a large body (the Sun). Einstein seems to postulate (more on the 'seems' below) that the Sun be represented by what would soon be recognized as an approximation to the Schwarzschild metric. He definitely postulates (!) that Mercury moves on a geodesic of said metric.[13] In a way, the problem of motion in GR is about the question of

tions (a special case of vacuum solutions) on the one hand and the vacuum approach to the problem of motion on the other hand is very promising indeed. I will have to postpone a detailed discussion to a later paper; it will include the problem of motion of a binary black hole, the black hole equivalent of the Sun-Mercury two-body system discussed below.

[13]For a careful analysis of Einstein's Mercury paper and how it rests on the Einstein-Besso manuscript see Earman and Janssen [1993], and Janssen's Editorial Note on the

whether this second postulate is really necessary.

If we now look at Einstein's Mercury paper and recall the kind of criticism that was launched against the vacuum approach to the problem of motion, we may find ourselves feeling puzzled. After all, the Schwarzschild metric is a solution to the vacuum field equations, and it has a singularity at its center.[14] If representing material bodies by singular metrics is so problematic, how does it come about that Einstein [1915] successfully predicted the perihelion motion of Mercury? Why is it not problematic to represent the Sun by the singular Schwarzschild metric?

The answer lies in denying the premise of the question. Einstein's treatment of the Sun-Mercury system should *not* be interpreted as involving him representing the Sun by (an approximation of) the Schwarzschild metric. We *know* that the Sun is a material body with non-vanishing mass-energy, and that it does not have a spacetime singularity at its center. What Einstein really does is to convert the two-body problem Sun-Mercury into a one-body problem, where one body (Mercury) is subject to an external gravitational field. It is the exterior gravitational field of the Sun, *not the Sun itself*, that is represented by the Schwarzschild metric. And that is enough to predict the perihelion of Mercury: we don't need to know what the Sun is made of or what happens in its interior; all that matters is the exterior gravitational field that Mercury is subject to.

Thus, worrying about the singularity at the center of the Schwarzschild metric just misses the point: we do not have to interpret the interior part of the Schwarzschild metric literally, at least not in this application.

In the following I shall argue that we should interpret the appearance of singularities in the Einstein-Grommer vacuum approach to the problem of motion in a similar vein.

3.2 The Einstein-Grommer vacuum approach to the problem of motion

The general scheme of the Einstein-Grommer approach proceeds as follows.[15]

Einstein-Besso manuscript in Vol. 4 of the Collected Papers of Albert Einstein (CPAE).

[14]For the history and interpretation of the Schwarzschild metric and its analytic extensions see Eisenstaedt [1989] and Bonnor [1992].

[15]The genesis of the Einstein-Grommer approach has been a bit of a mystery up to now, as pointed out by Kennefick [2005]. However, the work on the 15th volume of Einstein's collected papers has revealed the context and correspondence leading up to that paper,

1. Reformulate the vacuum Einstein equations in terms of a surface integral over a three-dimensional hyper-surface such that we can ask whether gravitational energy-momentum represented by the pseudo-tensor t^{τ}_{α} passes through the surface.[16]

2. Pick a curve that is supposed to represent the path of a material particle.

3. Impose the linear approximation according to which $g_{\mu\nu} = \eta_{\mu\nu} + \gamma_{\mu\nu}$, i.e. assume that, at least close to the curve, the metric deviates from Minkowski spacetime only slightly.

4. Realise that not all solutions to the linearized field equations will correspond to solutions of the non-linear field equations that the linearized field equations approximate. Argue that in the case where an 'equilibrium condition' for the energy-pseudo-tensor of the gravitational field holds, the $\gamma_{\mu\nu}$ of the linearized field equations *will* solve the full non-linear equations reformulated as a surface integral.[17]

5. Now split the $\gamma_{\mu\nu}$ in the immediate neighborhood of the particle into the 'inner metric $\bar{\gamma}_{\mu\nu}$ that the particle itself gives rise to and the 'outer metric $\bar{\bar{\gamma}}_{\mu\nu}$ that is due to other sources (or lack thereof). Observe that the 'outer metric' is entirely regular, even if extended to the point at which the material particle is supposed to be located.

6. Integrate the surface integral that is equivalent to the vacuum field equations 'around' the curve that is supposed to represent the path of a material particle. For the case where the integration surface is a sphere, the equilibrium condition for t^{τ}_{α} simplifies to $\frac{\partial \bar{\bar{\gamma}}_{44}}{\partial x_{\sigma}} = 0$.

---

and how it fits into Einstein's overall research program. It is a fascinating story; alas, it will have to wait for a separate paper.

[16]There has been a long debate on whether gravitational energy can be adequately represented by a pseudo-tensor; I will not be able to do it justice here. For some details see the introduction to Volume 8 CPAE for the debate between Einstein, Klein, Levi-Civita and Lorentz, for conceptual analysis Hoefer [2000] and especially Trautmann [1962].

[17]This step is very intricate and it would take me a few pages to do it justice. This point of the Einstein-Grommer paper has not been adressed by the literature at all (neither in physics nor in philosophy); I will argue elsewhere that it sheds new light on Einstein's later doubts as to whether the gravitational wave solutions of the linearized equations correspond to gravitational wave solutions in the full non-linear theory.

7. Conclude that the curve that represents the path of a material particle is a geodesic of the outer metric $\bar{\bar{\gamma}}_{\mu\nu}$.[18]

4 Interpreting the Einstein-Grommer approach to the problem of motion

The reader might think that the argument presented in the last section cannot be a faithful representation of the Einstein-Grommer approach; after all, where is the claim that the material particle is represented by a singularity, the reason the approach was dismissed by Earman and Tamir? Indeed, I have omitted that after step 5 of the argument Einstein and Grommer *do* say that one *could* assume that the inner metric $\bar{\gamma}_{\mu\nu}$ is given by what is effectively a three-dimensional counterpart of the Schwarzschild metric: it is spherically symmetric and has a singularity at the center. And yet, *Einstein and Grommer never use this assumption in their argument.* They call the material particle 'the singularity' all the time, but their argument does not depend on assuming *any* particular form for the inner metric, let alone one that is necessarily singular. As a matter of fact, they do not even mention a concrete candidate metric for the outer metric $\bar{\bar{\gamma}}_{\mu\nu}$; all they need is that $\gamma_{\mu\nu}$ is split into the inner metric $\bar{\gamma}_{\mu\nu}$ and the outer metric $\bar{\bar{\gamma}}_{\mu\nu}$ in such a way that $\bar{\bar{\gamma}}_{\mu\nu}$ is non-singular everywhere.

Note that this does not mean that we *know* that the inner metric $\bar{\gamma}_{\mu\nu}$ is non-singular. We don't know anything about the inner metric, for the argument is independent of $\bar{\gamma}_{\mu\nu}$ having any particular form, just like the derivation of Mercury's perihelion was independent of whether there is a singularity at the center of the Schwarzschild metric that represented the exterior field of the Sun.

With regard to the Sun-Mercury system I argued that we should not interpret the Schwarzschild metric as representing the Sun, but as representing its exterior gravitational field. The part of the Sun that is within the event horizon, including the singularity at the center, should not be taken

[18] Einstein and Grommer then go on to generalise this result to the 'non-stationary case', i.e. the case where it is not demanded that the external gravitational field, to which the particle is subject to, does not change in time. They conclude that in this case, too, the particle will move on a geodesic of the outer metric $\bar{\bar{\gamma}}_{\mu\nu}$ that is a solution to the field equations. For the following this generalisation does not make a difference; I will thus refer only to the stationary scenario described above.

as a representation of the *actual* interior of the Sun, but as a *placeholder* or a *blind spot* within the current description of the Sun-Mercury system: a docking station for a theoretical model of the Sun not included in Einstein's Sun-Mercury model.[19]

Likewise, we should interpret the inner metric $\bar{\gamma}_{\mu\nu}$ in the Einstein-Grommer approach as a placeholder for a representation of matter not included in the current theoretical approach. Sure, you *can* set $\bar{\gamma}_{\mu\nu}$ to be a Schwarzschild-like metric with a singularity at the center. But you don't have to do that to make the Einstein-Grommer argument work, and even if you do make that assumption, you should still take this particular inner metric with a singularity at its center as a placeholder for a representation or theory of matter not yet provided.[20]

But now wait a minute. You might have disliked the occurence of singularities as representations of particles, but at least the singularity (in lieu of a non-vanishing energy-momentum tensor) gave you an idea of *where* in spacetime the particle was supposed to be. True, Earman and Tamir rightly pointed out that the singularity is not actually part of spacetime, and so it can hardly serve to localize the particle in spacetime. Still, you might think that we're throwing the baby out with the bath water by not choosing any inner metric. After all, is it not the case then that the curve we have been focusing on is just *any* curve, without any reason to think of this curve as the curve of a material particle?[21]

Again, I think we can counter this criticism by comparing the Einstein-Grommer approach to Einstein's treatment of the Sun-Mercury system in

[19]Note that there are interior extensions of the Schwarzschild metric that model the interior of the Sun by solutions of the non-vacuum field equations (1), for example by an incompressible perfect fluid. See Bonnor [1992], section 5.

[20]If I had given more historical details, I could have, I believe, shown that Einstein himself saw the occurence of a singularity in the inner metric in exactly this way. This exegetical argument would have started with evidence that, from early on, he saw GR as a theory of the pure gravitational field without any constraints on what kinds of matter give rise to the gravitational field. Furthermore, I would have argued that even in the Einstein-Grommer paper he clearly forbids singularities *outside* of material particles (where the theory is supposed to give an adequate and deterministic representation of gravitational fields) but has no problem with them appearing *inside of* material systems, where the theory can provide at best phenomenological placeholders for a future 'proper' theory of matter anyhow. Thus, for Einstein energy-momentum tensors as alleged representatives of material systems were on a par with singularities: both were only placeholders for a proper theory of matter.

[21]I thank Jim Weatherall for putting this question to me.

Einstein [1915]. What Einstein did there was to assume that Mercury would move on *some* geodesic of the exterior gravitational field produced by the Sun. He calculated an approximation to the external gravitational field of a static, spherically symmetric and asymptotically flat body; this gravitational field he saw as represented by the connection components $\Gamma^{\nu}{}_{\mu\sigma}$ of a metric $g_{\mu\nu}$ which deviated only slightly from the flat Minkowski metric. He then inserted these gravitational field components $\Gamma^{\nu}{}_{\mu\sigma}$ into the geodesic equation (2). He showed that this law contained Newton's first law and Newton's second law with a gravitational potential giving rise to a force as a limiting case, and showed how the resulting Keplerian laws for orbits differ in his theory as compared to its Newtonian limit. In the end, he obtained that according to the new theory the perihelion ϵ of *any* geodesic orbit around the Sun is given by

$$\epsilon = 24\pi^3 \frac{a^2}{T^2 c^2 (1 - e^2)} \tag{5}$$

Here a denotes the length of the semimajor axis of the orbit in question, e its eccentricity, c the speed of light, and T the orbital period of the planet in question. Einstein then *takes the astronomically known values for Mercury*, plugs them into equation (5), and thereby predicts that Mercury's perihelion changes by 43" per century.

Note that there is *nothing* in the theoretical description that singles out any particular path as that of Mercury. There is no theoretical representation of Mercury, no model. All that is there is the assumption that Mercury will move on one of the geodesics of the affine connection determined by the spherically symmetric field of the Sun. A general equation that all possible geodesic orbits have to fulfil is derived. And then *external knowledge* is used to single out one of these orbits as that of Mercury. Einstein trusts that the astronomers have measured the orbital period, the semimajor axis and the eccentricity of Mercury correctly. It is this external knowledge, plugged into his theoretical model, which does not in itself contain a representation of Mercury or its path, that produces the prediction.

In many ways, the whole vacuum approach to the problem of motion is about the question as to whether in this kind of scenario we really have to assume the geodesic equation as the equation of motion of matter over and above the gravitational field equations. Indeed, let us look at the Sun-Mercury system within the 1927 Einstein-Grommer approach. The problem of motion, then, is the question whether Einstein really *had to* introduce the

gravitational field equations (to describe the exterior gravitational field of the Sun) *and* the geodesic equation (to describe the path of Mercury subject to this gravitational field) as separate assumptions.[22] Could he have only assumed the gravitational field equations and *derived* that Mercury moves on a geodesic of the exterior field of the Sun? My point is that, just like in Einstein's 1915 treatment, the 1927 Einstein-Grommer approach does not *need to* commit to a theoretical model that allows us to localise Mercury internally. It is fine to ask whether the exterior gravitational field around a given curve 'forces' that curve to be a geodesic. Just like in the 1915 treatment, Einstein and Grommer could then use *external knowledge* about whether that particular curve is actually the curve of a material object, or of Mercury in particular. No inner metric, no singularity to represent the material body, is actually needed.

Let us take a step back though, for there is an important difference between the structure of Einstein's 1915 treatment of Mercury on the one hand and the 1927 Einstein-Grommer approach on the other. In the Mercury case Einstein had assumed (!) that Mercury moves on a geodesic, i.e. a special kind of curve, and model-external knowledge about the period, eccentricity and semimajor axis of Mercury could then be used to determine which of the many geodesics of the Schwarzschild metric corresponded to the path of Mercury. But in the case of the Einstein-Grommer argument, what is in question is whether we can prove that the path of Mercury, say, is a geodesic. Thus, at first sight it looks as if while the 1915 argument only needed external knowledge to determine which geodesic is that of Mercury, appeal to external knowledge in the Einstein-Grommer case would have to determine a.) that this curve is a geodesic and b.) that it is the curve of a material body.

Einstein and Grommer did not aim to derive both a.) and b.). Instead, while Einstein in 1915 used external knowledge at the end of his argument, Einstein and Grommer in 1927 use it at the beginning. They start out by assuming that a given curve is the curve of a material particle, and then ask whether having a regular outer metric (which solves the vacuum field equations) around the curve means that the curve of this material particle,

[22]Interestingly, Einstein did not yet have the final gravitational field equations in the Mercury paper; he found them a week later, in his fourth paper of November 1915. However, the approximation of the Schwarzschild metric that he uses in the Mercury paper is an approximative solution of both the field equations from the Mercury paper, and of the final Einstein field equations.

given the further conditions summarized in section 3.2, *must be* a geodesic. Rather than finishing the argument by appeal to external knowledge (as in Einstein 1915), the Einstein-Grommer argument starts with an appeal to external knowledge, which singles out a particular curve as that of a material body.[23]

Either way, both in Einstein's 1915 treatment and in the Einstein-Grommer approach there is no reason to interpret the singularity (appearing in the Schwarzschild metric or the inner metric, respectively) literally. In both cases, the singularity should be interpreted to signify a placeholder or a blind spot of the theoretical treatment, rather than something that should be interpreted literally, as referring and approximately true. Indeed, both Einstein's 1915 treatment of the Sun-Mercury system and Einstein's and Grommer's treatment of an arbitrary material particle subject to an external gravitational field work just as well if, in the former case, no interior metric (to describe the interior of the Sun) or, in the latter case, no inner metric (to represent the location of the particle on the curve), is ever specified.

5 Conclusion

I started out by saying that whether we are realists or antirealists, we should aim for a careful interpretation, rather than a literal interpretation, of the scientific theory that we want to be realists or anti-realists about. As a case study, I argued that the vacuum approach to the problem of motion in GR, and the Einstein-Grommer approach in particular, is far more sensible and promising if we interpret the singularities *not as representing* material bodies but as *placeholders* for a representation of material bodies that is not included in the model. Indeed, I argued that the approach does not even need the

[23]There is a further disanalogy between Einstein's 1915 derivation of the perihelion of Mercury and the Einstein-Grommer argument of 1927. In the former the choice of (an approximation) the Schwarzschild metric to represent the exterior gravitational field of the Sun does important work in the derivation of Mercury's perihelion. In the Einstein-Grommer approach, no choice of a concrete outer metric is necessary to derive that the curve of the particle which is surrounded by the outer metric must be a geodesic. The reason for this difference is that the Einstein-Grommer approach aims to be more general; it only aims to derive *that* a material body moves on *some* geodesic of the outer metric. However, note that it is not the case that any outer metric is allowed by the approach: the class of outer metrics that the approach can work with is heavily constrained by steps 2 and 3 of the Einstein-Grommer argument (see section 3.2).

introduction of singularities to represent material bodies; their introduction does not do any work in answering the question at hand.[24]

Given that in their paper Einstein and Grommer seem to take the singularities as representing material bodies, one might wonder whether this allegedly more careful interpretation does not fall prey to the criticism that the careful interpreter presumes to understand the theory/formalism in question better than its originators. This might seem at odds with the realist tenet of taking scientists and science 'seriously'. I do indeed think that putting the Einstein-Grommer paper into its proper historical context by analysing Einstein's correspondence leading up to the paper and by relating it to his overarching research project at the time *would* convincingly show that he subscribed to something very much like the 'placeholder interpretation' I defended above. Showing this in detail will have to wait for a much longer paper, and I do not ask the reader to just take my word for it. So let us say, for the sake of the argument, that Einstein and Grommer did indeed intend the singularities as representatives of material objects in a rather straightforward way. I believe that we should not take *their* word for it either. And neither did Einstein. Just a few years after the Einstein-Grommer paper, in his famed 1933 Spencer lectures at the University of Oxford, Einstein told us in his opening words: "If you wish to learn from the theoretical physicist anything about the methods which he uses, I would give you the following advice: Don't listen to his words, examine his achievements."[25]

In philosophy of science, I believe there is no better way of examining a scientist's achievements than by looking for the best possible interpretation

[24]The argument that we should thus not see a realist as comitted to being a realist *about* the singularities appearing in the Einstein-Grommer paper resonates well with selective or posit realism as introduced by Vickers [2013]. The idea there is that we should only be realists with respect to components of a prediction that 'fuel the success' of the prediction, i.e. that are indispensable in the derivation of what is predicted. Using Vickers' distinction the introduction of a singular inner metric in the Einstein-Grommer approach is an idle rather than a working posit. However, note that the call for careful rather than literal interpretations with which I started is independent of / complementary to aiming for identification of the idle posits in a derivation. For *even if* we had found that the introduction of the singular inner metric did do work in the derivation of geodesic motion could we have argued (with less force) that the singularity should be interpreted as a placeholder for a future theory of matter, as a temporary measure within an effective theory, and thus not as something that we should interpret as possesing as much 'reality' or 'referring power' as the regular outer metric governed by the field equations.

[25]See Einstein [1934], and van Dongen [2010] for a detailed analysis of the text.

of his or her theories. To do that, we have to not just listen to the words of the scientist who created or discovered it; we have to see what the theory *does* in practice, how it is *used*; which of its parts really do the work.

Acknowledgments

I would like to thank my colleagues at Caltech and at the Einstein Papers Project for many discussions about the problem of motion and the Einstein-Grommer approach in particular. Thanks are due especially to Diana Kormos-Buchwald, Frederick Eberhardt, and Daniel Kennefick. I would also like to thank audiences at Caltech, Oxford, Irvine, the BSPS 2016 conference in Cardiff, and at the 8th Quadrennial Pittsburgh Fellows conference in Lund, Sweden for many helpful discussions on the topic. I would like to thank especially Sam Fletcher, David Malament and Jim Weatherall for carefully reading earlier versions of this paper, and for the extremely helpful comments they gave me. Finally, I would like to thank Dana Tulodziecki for pointing my attention to the link between posit realism and what I was saying in this paper.

References

Abbott, B., Abbott, R., Abbott, T., Abernathy, M., Acernese, F., Ackley, K., Adams, C., Adams, T., Addesso, P., Adhikari, R. et al. [2016], 'Observation of gravitational waves from a binary black hole merger', *Physical Review Letters* **116**(6), 061102.

Bonnor, W. [1992], 'Physical interpretation of vacuum solutions of einstein's equations. part i. time-independent solutions', *General relativity and Gravitation* **24**(5), 551–574.

Brown, H. R. [2007], *Physical Relativity. Space-time structure from a dynamical perspective*, Oxford University Press, USA.

Curiel, E. [1999], 'The analysis of singular spacetimes', *Philosophy of Science* pp. S119–S145.

Curiel, E. [Forthcoming], A primer on energy conditions, *in* D. Lehmkuhl,

G. Schiemann and E. Scholz, eds, 'Towards a Theory of Spacetime Theories', Einstein Studies, Birkhäuser.

Earman, J. [1995], *Bangs, crunches, whimpers, and shrieks: Singularities and acausalities in relativistic spacetimes*, Oxford University Press, USA.

Earman, J. and Eisenstaedt, J. [1999], 'Einstein and singularities', *Studies in History and Philosophy of Modern Physics* **30**(2), 185–235.

Earman, J. and Janssen, M. [1993], 'Einstein's explanation of the motion of mercury's perihelion', *Einstein Studies* pp. 129–172.

Ehlers, J. and Geroch, R. [2004], 'Equation of motion of small bodies in relativity', *Annals of Phyiscs* **309**, 232–236.

Einstein, A. [1915], 'Erklärung der perihelbewegung des merkur aus der allgemeinen relativitätstheorie', *Königliche Preussische Akademie der Wisenschaften (Berlin)* .

Einstein, A. [1922], *Vier Vorlesungen über Relativitätstheorie gehalten im Mai 1921 an der Universität Princeton*, F. Vieweg. Reprinted as Vol.7, Doc. 71 CPAE; and in various editions as "The Meaning of Relativity" by Princeton University Press.

Einstein, A. [1934], 'On the method of theoretical physics', *Philosophy of science* **1**(2), 163–169.

Einstein, A. [1936], 'Physics and reality', *Journal of the Franklin Institute* **221**, 349–382. Reprinted in A. Einstein (1976) Ideas and Opinions (New York: Dell Publishers), pp. 283-315.

Einstein, A. and Grommer, J. [1927], 'Allgemeine Relativitätstheorie und Bewegungsgesetz', *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften, Physikalisch-Mathematische Klasse* pp. 2–13.

Eisenstaedt, J. [1989], The early interpretation of the schwarzschild solution, *in* D. H. a. J. Stachel, ed., 'Einstein and the History of General Relativity', Birkhäuser, pp. 1–213.

Geroch, R. and Jang, P. [1975], 'Motion of a body in general relativity', *Journal of Mathematical Physics* **16**, 65–67.

Havas, P. [1989], The early history of the "problem of motion" in general relativity, *in* 'Einstein and the History of General Relativity', Vol. 1, pp. 234–276.

Hoefer, C. [2000], 'Energy conservation in gtr', *Studies in History and Philosophy of Modern Physics* **31**.

Kennefick, D. [2005], 'Einstein and the problem of motion: a small clue', *The universe of general relativity* pp. 109–124.

Lehmkuhl, D. [2014], 'Why einstein did not believe that general relativity geometrizes gravity', *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* **46**, 316–326.

Malament, D. B. [2012], A remark about the "geodesic principle" in general relativity, *in* 'Analysis and Interpretation in the Exact Sciences', Springer, pp. 245–252.

Pickering, A. [1999], *Constructing quarks: A sociological history of particle physics*, University of Chicago Press.

Saunders, S., Barrett, J., Kent, A. and Wallace, D. [2010], *Many worlds? Everett, quantum theory, & reality*, OUP Oxford.

Tamir, M. [2012], 'Proving the principle: Taking geodesic dynamics too seriously in Einstein's theory', *Studies In History and Philosophy of Modern Physics* **43**(2), 137–154.

Torretti, R. [1996], *Relativity and geometry*, Dover Publications.

Trautmann, A. [1962], Conservation laws in general relativity, *in* L. Witten, ed., 'Gravitation: An Introduction to Current Research', John Wiley and Sons.

Tupper, B. [1981], 'The equivalence of electromagnetic fields and viscous fluids in general relativity', *Journal of Mathematical Physics* **22**(11), 2666–2673.

Tupper, B. [1982], 'The equivalence of perfect fluid space-times and magnetohydrodynamic space-times in general relativity', *General Relativity and Gravitation* **15**(1).

Tupper, B. [1983], 'The equivalence of perfect fluid space-times and viscous magnetohydrodynamic space-times in general relativity', *General Relativity and Gravitation* **15**(9).

van Dongen, J. [2010], *Einstein's Unification*, Cambridge University Press, Cambridge.

Van Fraassen, B. C. [1980], *The scientific image*, Oxford University Press.

Vickers, P. [2013], 'A confrontation of convergent realism', *Philosophy of Science* **80**(2), 189–211.

Weatherall, J. O. [2011], 'On the status of the geodesic principle in newtonian and relativistic physics', *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics* **42**(4), 276 – 281.
URL: *http://www.sciencedirect.com/science/article/pii/S1355219811000566*

Weatherall, J. O. [Forthcoming], Inertial motion, explanation, and the foundations of classical spacetime theories, *in* D. Lehmkuhl, G. Schiemann and E. Scholz, eds, 'Towards a Theory of Spacetime Theories', Birkhäuser.

Holism, or the Erosion of Modularity – a Methodological Challenge for Validation

Draft to be presented at PSA 2016

Johannes Lenhard, Bielefeld University

abstract

Modularity is a key concept in building and evaluating complex simulation models. My main claim is that in simulation modeling modularity degenerates for systematic methodological reasons. Consequently, it is hard, if not impossible, to accessing how representational (inner mathematical) structure and dynamical properties of a model are related. The resulting problem for validating models is one of holism.

The argument will proceed by analyzing the techniques of parameterization, tuning, and kludging. They are – to a certain extent – inevitable when building complex simulation models, but corrode modularity. As a result, the common account of validating simulations faces a major problem and testing the dynamical behavior of simulation models becomes all the more important. Finally, I will ask in what circumstances this might be sufficient for model validation.

1. Introduction

For the moment, imagine a scene at a car racing track. The air smells after gasoline. The pilot of the F1 racing car has just steered into his box and is peeling himself out of the straight cockpit. He puts off his helmet, shakes his sweaty hair, and then his eyes make contact to the technical director with a mixture of anger, despair, and helplessness. The engine had not worked as it should, and for a known reason: the software. However, the team had not been successful in attributing the miserable performance to a particular parameter setting. The machine and the software interacted in unforeseen and intricate ways. This explains the exchange of glances between pilot and technical director. The software's internal interactions and interfaces proved to be so complicated that the team had not been able to localize an error or a bug, rather remained

suspicious that some complex interaction of seemingly innocent assumptions or parameter settings was leading to the insufficient performance.

The story happened in fact[1] and it is remarkable since it displays how invasive computational modeling is into areas that smell most analogous. I reported this short piece for another reason, however, namely because the situation is typical for complex computational and simulation models. Validation procedures, while counting on modularity, run against a problem of holism.

Both concepts, modularity and holism, are notions at the fringe of philosophical terminology. Modularity is used in many guises and is not a particularly philosophical notion. It features prominently in the context of complex design, planning, and building – from architecture to software. Modularity stands for first breaking down complicated tasks into small and well-defined sub-tasks and then re-assembling the original global task with a well-defined series of steps. It can be argued that modularity is the key pillar on which various rational treatments of complexity rest – from architecture to software engineering.

Holism is a philosophical term to a somewhat higher degree and is covered in recent compendia. The Stanford Encyclopedia, for instance, includes (sub-)entries on methodological, metaphysical, relational, or meaning holism. Holism generically states that the whole is greater than the sum of its parts, meaning that the parts of a whole are in intimate interconnection, such that they cannot exist independently of the whole, or cannot be understood without reference to the whole. Especially W. V. O. Quine has made the concept popular, not only in philosophy of language, but also in philosophy of science, where one speaks of the so-called Duhem-Quine thesis. This thesis is based on the insight that one cannot test a single hypothesis in isolation, but that any such test depends on "auxiliary" theories or hypotheses, for example how the measurement instruments work. Thus any test addresses a whole ensemble of theories and hypotheses.

Lenhard and Winsberg (2010) have discussed the problem of confirmation holism in the context of validating complex climate models. They argued that "due to interactivity, modularity does not break down a complex system into separately manageable pieces." (2010, 256) In a sense, I want to pick up on this work, but put the thesis into a much more general context, i.e. pointing

[1] In spring 2014, the Red Bull team experienced a crisis due to recalcitrant problems with the Renault engine, due to a partial software update.

out a dilemma that is built on the tension between modularity and holism and that occurs quite generally in simulation modeling. The potential philosophical novelty is debated controversially in philosophy of science, for instance Humphreys (2009) vs. Frigg and Reiss (2009). The latter authors deny novelty, but concede issues of holism might be an exception. My paper confirms that holism is a key concept when reasoning about simulation. (I see more reasons for philosophical novelty, though.)

My main claim is the following: According to the rational picture of design, modularity is a key concept in building and evaluating complex models. In simulation modeling, however, modularity erodes for systematic methodological reasons. Moreover, the very condition for success of simulation undermines the most basic pillar of rational design. Thus the resulting problem for validating models is one of (confirmation) holism.

Section 2 discusses modularity and its central role for the so-called rational picture of design. Herbert Simon's highly influential parable of the watchmakers will feature prominently. It paradigmatically captures complex systems as a sort of large clockwork mechanism. This perspective suggests the computer would enlarge the tractability of complex systems due to its vast capacity for handling (algorithmic) mechanisms. Complex simulations then would appear as the electronic incarnation of a gigantic assembly of cogwheels. This viewpoint is misleading, I will argue. Instead, I want to emphasize the dis-analogy to how simulation models work. The methodology of building complex simulation models thwarts modularity in systematic ways. Simulation is based on an iterative and exploratory mode of modeling that leads to a sort of *holism that erodes modularity*.

I will present two arguments for the erosion claim, one from parameterization and tuning (section 3), the other from klu(d)ging (section 4). Both are, in practice, part-and-parcel of simulation modeling and both make modularity erode. The paper will conclude by drawing lessons about the limits of validation (section 5). Most accounts of validation require, if often not explicitly, modularity and are incompatible with holism. In contrast, the exploratory and iterative mode of modeling restricts validation, at least to a certain extent, to testing (global) predictive virtues. This observation shakes the rational (clockwork) picture of design and of the computer.

2. The rational picture

The design of complex systems has a long tradition in architecture and engineering. At the same time, it has not been much covered in literature, because design was conceived as a matter for experienced craftsmanship rather than analytical investigation. The work of Pahl and Beitz (1984, plus revised editions 1996, 2007) gives a relatively recent account of design in engineering. A second, related source for reasoning about design is the design of complex computer systems. Here, one can find more explicit accounts, since the computer led to complex systems much faster than any tradition of craftsmanship could grow. A widely read example is Herbert Simon's "Sciences of the Artificial" (1969). Still up to today, techniques of high-level languages, object-oriented programming, etc. make the practice of design change on a fast scale.

One original contributor to this discussion is Frederic Brooks, software and computer expert (and former manager at IBM) and also hobby architect. In his 2010 monograph "The Design of Design", he describes the rational model of design that is widely significant, though it is much more often adopted in practice than explicitly formulated in theoretical literature. The rational picture starts with assuming an overview of all options at hand. According to Simon, for instance, the theory of design is the general theory of search through large combinatorial spaces (Simon 1969, 54). The rational model then presupposes a utility function and a design tree, which are spanning the space of possible designs. Brooks rightly points out that these are normally not at hand. Nevertheless, design is conceived as a systematic step-by-step process. Pahl and Beitz aim at detailing these steps in their rational order. Also, Simon presupposes the rational model, arguably motivated by making design feasible for artificial intelligence (see Brooks 2010, 16). Wynston Royce, to give another example, introduced the "waterfall model" for software design (1970). Royce was writing about managing the development of large software systems and the waterfall model consisted in following a hierarchy ("downward"), admitting to iterate steps on one layer, but not with much earlier ("upward") phases of the design process. Although Royce actually saw the waterfall model as a straw man, it was cited positively as paradigm of software development (cf. Brooks on this point).

Some hierarchical order is a key element of the rational picture of design and presumes modularity. Let me illustrate this point. Consider first a simple brick wall. It consists of a multitude of modules, each with certain form and static properties. These are combined into

potentially very large structures. It is a strikingly simple example, because all modules (bricks) are similar.

A more complicated, though closely related, example is the one depicted in figure 1 where an auxiliary building of Bielefeld University is put together from container modules.

Figure 1: A part of Bielefeld University is built from container modules.

These examples illustrate how deeply ingrained modularity is in our way of building (larger) objects. Figure 2 displays a standard picture for designing and developing complex (software) systems.

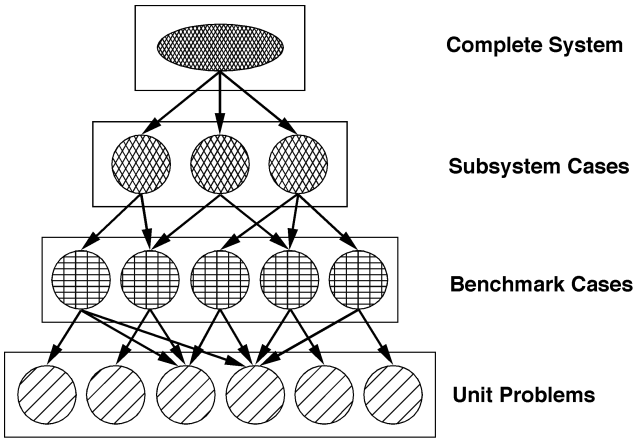


Figure 2: Generic architecture of complex software, from the AIAA Guide for the Verification and Validation of Computational Fluid Dynamics Simulations (1998). Modules of one layer might be used by different modules on a higher layer.

Some complex overall task is split up into modules that can be tackled independently and by different teams. The hierarchical structure shall ensure the modules can be integrated to make up the original complex system. Modularity not only plays a key role when designing and building complex systems, it also is of crucial importance when taking account of the system. Validation is usually conceived in the very same modular structure: independently validated modules are put together in a controlled way for making up a validated bigger system. The standard account of how computational models are verified and validated gives very rigorous guidelines that are all based on the systematic realization of modularity (Oberkampf and Roy 2010, see also Fillion 2017). In short, modularity is key for designing as well as for validating complex systems.

This observation is paradigmatically expressed in Simon's parable of the two watchmakers. You find it in Simon's 1962 paper "The Architecture of Complexity" that has become a chapter in his immensely influential "The Sciences of the Artificial" (Simon 1969). There, Simon investigates the structure of complex systems. The stable structures, so Simon argues, are the hierarchical ones. He expressed his idea by telling the parable of the two watchmakers named Hora and Tempus (1969, 90-92). P. Agre describes the setting with the following words:

"According to this story, both watchmakers were equally skilled, but only one of them, Hora, prospered. The difference between them lay in the design of their watches. Each design involved 1000 elementary components, but the similarity ended there. Tempus' watches were not hierarchical; they were assembled one component at a time. Hora's watches, by contrast, were organized into hierarchical subassemblies whose "span" was ten. He would combine ten elementary components into small subassemblies, and then he would combine ten subassemblies into larger subassemblies, and these in turn could be combined to make a complete watch." (Agre 2003)

Since Hora takes additional steps for building modules, Tempus' watches need less time for assembly. However, it was Tempus' business that did not thrive, because of an additional condition not yet mentioned, namely some kind of noise. From time to time the telephone rings and whenever one of the watchmakers answers the call, all cogwheels and little screws fall apart and he has to re-start the assembly. While Tempus had to start from scratch, Hora could keep all finished modules and work from there. In the presence of noise, so the lesson goes, the modular

strategy is by far superior. Agre summarizes that modularity, he speaks of the functional role of components, comes out as a necessary element when designing complex systems:

"For working engineers, hierarchy is not mainly a guarantee that subassemblies will remain intact when the phone rings. Rather, hierarchy simplifies the process of design cognitively by allowing the functional role of subassemblies to be articulated in a meaningful way in terms of their contribution to the function of the whole. Hierarchy allows subassemblies to be modified somewhat independently of one another, and it enables them to be assembled into new and potentially unexpected configurations when the need arises. A system whose overall functioning cannot be predicted from the functionality of its components is not generally considered to be well-engineered." (Agre 2003)

Now, the story works with rather particular examples insofar as watches exemplify complicated mechanical devices. The universe as a giant clockwork has been a common metaphor since the seventeenth century. Presumably, Simon was aware the clockwork picture is limited and he even mentioned that complicated interactions could lead to a sort of pragmatic holism.[2] Nonetheless, the hierarchical order is established by the interaction of self-contained modules.

There is an obvious limit to the watchmaker picture, namely systems have to remain manageable by human beings (watchmakers). There are many systems of practical interest that are too complex – from the earth's climate to the aerodynamics of an airfoil. Computer models open up a new path here, since simulation models might contain a wealth of algorithmic steps far beyond what can be conceived in a clockwork picture.[3] From this point of view, the computer appears as a kind of amplifier that helps to revitalize the rational picture. Do we have to look at simulation models as a sort of gigantic clockworks? In the following, I will argue that this viewpoint is seriously misleading. Simulation models are different from watches in important ways and I

[2] This kind of holism hence can occur even when modules are "independently validated", since these modules when connected together could interact with each other in unpredicted ways. This is a strictly weaker form of holism than the one I am going to discuss.
[3] Charles Babbage had designed his famous „Analytical Engine" as a *mechanistic* computer. Tellingly, it did encounter serious problems exactly because of the mechanical limitations of its construction.

want to focus on the dis-analogy.[4] Finally, we will learn from the investigation of simulation models about our picture of rationality.

3. Erosion of modularity 1: Parameterization and tuning

In stark contrast to the cogwheel picture of the computer, the methodology of simulation modeling erodes modularity in systematic ways. I want to discuss two separate though related aspects, firstly, parameterization and tuning and, secondly, kluging (also called kludging). Both are, for different reasons, part-and-parcel of simulation modeling; and both make modularity of models erode. Let us investigate them in turn and develop two arguments for erosion.

Parameterization and tuning are key elements of simulation modeling that stretch the realm of tractable subject matter much beyond what is covered by theory. Furthermore, simulation models can make predictions even in fields that *are* covered by well-accepted theory only with the help of parameterization and tuning. In this sense, the latter are success conditions for simulations.

Before we start with discussing an example, let me add a few words about terminology. There are different expressions that specify what is done with parameters. The four most common ones are (in alphabetical order): adaptation, adjustment, calibration, and tuning. These notions describe very similar activities, but also valuate differently what parameters are good for. Calibration is commonly used in the context of preparing an instrument, like calibrating a scale one time for using it very often in a reliable way. Tuning has a more pejorative tone, like achieving a fit with artificial measures, or fitting to a particular case. Adaptation and adjustment have more neutral meanings.

Atmospheric circulation is a typical example. It is modeled on the basis of accepted theory (fluid dynamics, thermodynamics, motion) on a grand scale. Climate scientists call this the "dynamical core" of their models and there is more or less consensus about this part. Although the employed theory is part of physics, climate scientists mean a different part of their models when they speak of "the physics". It includes all the processes that are not completely specified from the dynamical core. These processes include convection schemes, cloud dynamics, and many more.

[4] There are several dis-analogies. One I am not discussing is that clockworks lack multi-functionality.

The "physics" is where different models differ and the physics is what modeling centers regard as their achievements and try to maintain even if their models change into the next generation.

The physics acts like a specifying supplement to the grand scale dynamics. It is based on modeling assumptions, say which sub-processes are important in convection, what should be resolved in the model and what should be treated via a parameterization scheme. Often, such processes are not known in full detail, and some aspects (at least) depend on what happens on a sub-grid scale. The dynamics of clouds, for instance, depends on a staggering span of very small (molecular) scales and much larger scales of many kilometers. Hence even if the laws that guide these processes would be known, they could not be treated explicitly in the simulation model. Modeling the physics has to bring in parameterization schemes.[5]

How does moisture transport, for example, work? Rather than trying to investigate into the molecular details of how water vapor is entrained into air, scientists use a parameter, or a scheme of parameters, that controls moisture uptake so that known observations are met. Often, such parameters do not have a direct physical interpretation, nor do they need one, like when a parameter stands for a mixture of processes not resolved in the model. The important property rather is that they make the parameterization scheme flexible, so that the parameters of such a scheme can be changed in a way that makes the properties of the scheme (in terms of climate dynamics) match some known data or reference points.

From this rather straightforward observation follows an important fact. A parameterization, including assignments of parameter values, makes sense only in the context of the larger model. Observational data are not compared to the parameterization in isolation. The Fourth Assessment Report of the IPCC acknowledges the point that "parameterizations have to be understood in the context of their host models" (Solomon et al. 2007, 8.2.1.3)

The question of whether the parameter value that controls moisture uptake (in our oversimplified example) is adequate can be answered only by examining how the entire parameterization behaves and, moreover, how the parameter value in the parameterization in the larger simulation model behaves. Answering such questions would require, for instance, looking at more global properties like mean cloud cover in tropical regions, or the amount of rain in some area. Briefly

[5] Parameterization schemes and their more or less autonomous status are discussed in the literature, cf. Parker 2013, Smith 2002, or Gramelsberger and Feichter 2011.

stated, parameterization is a key component of climate modeling and tuning is part-and-parcel of parameterization.[6]

It is important to note that tuning one parameter takes the values of other parameters as given, be they parameters from the same scheme, or be they parts of other schemes that are part of the model. A particular parameter value (controlling moisture uptake) is judged according to the results it yields for the overall behavior (like cloud cover). In other words, tuning is a local activity that is oriented at global behavior. Researchers might try to optimize parameter values simultaneously, but for reasons of computational complexity, this is possible only with a rather small subset of all parameters. A related issue is statistical regression methods that might be caught up in a local optimum. In climate modeling, skill and experience remain to be important for tuning (or adjustment).

Furthermore, tuning parameters is not only oriented at the global model performance, it tends to blur the local behavior. This is because every model will be importantly imperfect, since it contains technical errors, works with insufficient knowledge, etc. – which is just the normal case in scientific practice. Now, tuning a parameter according to the overall behavior of the model then means that the errors, gaps, and bugs get compensated against each other (if in an opaque way). Mauritsen et al. (2012) have pointed this out in their pioneering paper about tuning in climate modeling.

In climate models, cloud parameterizations play an important role, because they influence key statistics of the climate and, at the same time, cover major (remaining) uncertainties about how an adequate model should look like. Typically, such a parameterization scheme includes more than two dozens of parameters; most of them do not carry a clear physical interpretation. The simulation then is based on the balance of these parameters in the context of the overall model (including other parameterizations). Over the process of adjusting the parameters, these schemes become inevitably convoluted. I leave aside that models of atmosphere and oceans get coupled, which arguably aggravates the problem.

[6] The studies of so-called perturbed physics ensembles convincingly showed that crucial properties of the simulation models hinge on exactly how parameter values are assigned (Stainforth et al. 2007).

Tuning is inevitable, part-and-parcel of simulation modeling methodology. It poses great challenges, like finding a good parameterization scheme for cloud dynamics, which is a recent area of intense research in meteorology. But when is a parameterization scheme a good one? On the one side, a scheme is sound when it is theoretically well motivated, on the other side, the key property of a parameterization scheme is its adaptability. Both criteria do not point into the same direction. There is, therefore, no optimum; finding a balance is still considered an art. I suspect that the widespread reluctance against publishing about practices of adjusting parameters comes from reservations against aspects that call for experience and art rather than theory and rigorous methodology.

I want to maintain that nothing in the above argumentation is particular to climate. Climate modeling is just one example out of many. The point holds for simulation modeling quite generally. Admittedly, climate might be a somewhat peculiar case, because it is placed in a political context where some discussions seem to require that only ingredients of proven physical justification and realistic interpretation are admitted. Arguably, this expectation might motivate using the pejorative term of tuning. This reservation, however, ignores the very methodology of simulation modeling. Adjusting parameters is by no means particular to climate modeling, nor is it confined to areas where knowledge is weak.

Another example will document this. Adjusting parameters is also occurring thermodynamics, an area of physics with very high theoretical reputation. The ideal gas equation is even taught in schools, it is a so-called equation of state (EoS) that describes how pressure and temperature depend on each other. However, actually using thermodynamics requires to work with less idealized equations of state than the ideal gas equation. More complicated equations of state find wide applications also in chemical engineering. They are typically very specific for certain substances and require extensive adjustment of parameters as Hasse and Lenhard (2017) describe and analyze. Clearly, being able to process specific adjustment strategies that are based on parameterization schemes is a crucial success condition. Simulation methods have made applicable thermodynamics in many areas of practical relevance, exactly because equations of state are tailored to particular cases of interest via adjusting parameters.

One further example is from quantum chemistry, namely the so-called density functional theory (DFT), a theory developed in the 1960s that won the Nobel prize in 1998. Density functionals

capture the information of the Schroedinger equation, but are much more computationally tractable. However, only many-parameter functionals brought success in chemistry. The more tractable functionals with few parameters worked only in simpler cases of crystallography, but were unable to yield predictions accurate enough to be of chemical interest. Arguably, being able to include and adjust more parameters has been the crucial condition that had to be satisfied before DFT could gain traction in computational quantum chemistry, which happened around 1990. This traction, however, is truly impressive. DFT is by now the most widely used theory in scientific practice, see Lenhard (2014) for a more detailed account of DFT and the development of computational chemistry.

Whereas the adjustment of parameters – to use the more neutral terminology – is pivotal for matching given data, i.e. for predictive success, this very success condition also entails a serious disadvantage.[7] Complicated schemes of adjusted parameters might block theoretical progress. In our climate case, any new cloud parameterization that intends to work with a more thorough theoretical understanding has to be developed for many years and then has to compete with a well-tuned forerunner. Again, this kind of problem is more general. In quantum chemistry, many-parameter adaptations of density functionals have brought great predictive success but at the same time render the rational re-construction of why such success occurs hard, if not impossible (Perdew et al. 2005, discussed in Lenhard 2014). The situation in thermodynamics is similar, cf. Hasse and Lenhard (2017).

Let us take stock regarding the first argument for the erosion of modularity. Tuning, or adjusting, parameters is not merely an *ad hoc* procedure to smoothen a model, rather it is a pivotal component for simulation modeling. Tuning convolutes heterogeneous parts that do not have a common theoretical basis. Tuning proceeds holistically, on basis of global model behavior. How particular parts function often remains opaque. By interweaving local and global considerations, and by convoluting the interdependence of various parameter choices, tuning destructs modularity.

Looking back to Simon's clockmaker story, we see that its basic setting does not match the situation in a fundamental way. The perfect cogwheel picture is misleading, because it presupposes a clear identification of mechanisms and their interactions. In our examples, we saw

[7] There are other dangers, like over-fitting, that I leave aside.

that building a simulation model, different from building a clockwork, cannot proceed top-down. Moreover, different modules and their interfaces get convoluted during the processes of mutual adaptation.

4. Erosion of modularity 2: kluging

The second argument for the erosion of modularity approaches the matter from a different angle, namely from a certain practice in developing software known as kluging (also spelled kludging)[8]. "Kluge" is a term from colloquial language that became a term in computer slang. I remember when back in my childhood our family and another, befriended one drove towards holidays in two cars. In the middle of the night, while crossing the Alps, the exhaust pipe of our friends before us broke, creating a shower of sparks where the pipe met the asphalt. There was no chance of getting the exhaust pipe repaired, but the father did not hesitate long and used his necktie to fix it provisionally.

The necktie worked as a kluge, which is in the words of Wikipedia "a workaround or quick-and-dirty solution that is clumsy, inelegant, difficult to extend and hard to maintain, yet an effective and quick solution to a problem." The notion has been incorporated and become popular in the language of software programming and is closely related to the notion of bricolage.

Andy Clark, for instance, stresses the important role played by kluges in complex modular computer modeling. For him, a kluge is "an inelegant, 'botched together' piece of program; something functional but somehow messy and unsatisfying", it is—Clark refers to Sloman—"a piece of program or machinery which works up to a point but is very complex, unprincipled in its design, ill-understood, hard to prove complete or sound and therefore having unknown limitations, and hard to maintain or extend". (Clark 1987, 278)

Kluges carried forward their way from programmers' colloquial language into the body of philosophy guided by scholars like Clark and Wimsatt who are inspired both by computer

[8] Both spellings „kluge" and „kludge" are used. There is not even agreement of how to pronounce the word. In a way, that fits to the very concept. I will use "kluge", but will not change the habits of other authors cited with "kludge".

modeling and evolutionary theory.[9] The important point in our present context is that kluges may function for a whole system, i.e. for the performance of the entire simulation model, whereas it has no meaning in relation to the submodels and modules: "what is a kludge considered as an item designed to fulfill a certain role in a large system, may be no kludge at all when viewed as an item designed to fulfill a somewhat different role in a smaller system." (Clark 1987, 279)

Since kluging stems from colloquial language and is not seen as a good practice anyway, examples cannot be found easily in published scientific literature. This observation notwithstanding, kluging is a widely occurring phenomenon. Let me give an example that I know from visiting an engineering laboratory. There, researchers (chemical process engineers) are working with simulation models of an absorption column, the large steel structures in which reactions take place under controlled conditions. The scientific details do not matter here, since the point is that the engineers build their model on the basis of a couple of already existing modules, including proprietary software that they integrate into their simulation without having access to the code. Moreover, it is common knowledge in the community that this (unknown) code is of poor quality. Because of programming errors and because of ill-maintained interfaces, using this software package requires modifications on the part of the remaining code outside the package. These modifications are there for no good theoretical reason, albeit for good practical reasons. They make the overall simulation run as expected (in known cases); and they allow working with existing software. The modifications thus are typical kluges.

Again, kluging occurs in virtually every site where large software programs are built. Simulation models hence are a prime instance, especially when the modeling steps of one group build on the results (models, software packages) of other groups. One common phenomenon is the increasing importance of "exception handling", i.e. of finding effective repairs when the software, or the model, performs in unanticipated and undesired ways. In this situation, the software might include a bug that is invisible (does not affect results) most of the time, but becomes effective under particular conditions. Often extensive testing is needed for finding out about unwanted behavior that occurs in rare and particular situations that are conceived of as "exceptions", indicating that researchers do not aim at a major reconstruction, but at a local repair,

[9] The cluster of notions like bricolage and kluging common in software programming and biological evolution would demand a separate investigation. See, as a teaser, Francois Jacob's account of evolution as bricolage (1994).

counteracting this particular exception. Exception handling can be part of a sound design process, but increased use of exception handling is symptomatic of excessive kluging.

Presumably all readers who ever contributed to a large software program know about experiences of this kind. It is commonly accepted that the more comprehensive a piece of software gets, the more energy for exception handling new releases will require. Operating systems of computers, for example, often receive weekly patches. Many scientists who work with simulations are in a similar situation, though not obviously so.

If, for instance, meteorologists want to work on, say, hurricanes, they will likely take a meso-scale (multi-purpose) atmospheric model from the shelf of some trusted modeling center and add specifications and parameterizations relevant for hurricanes. Typically, they will not know in exactly what respects the model had been tuned, and also lack much other knowledge about strengths and weaknesses of this particular model. Consequently, when preparing their hurricane modules, they will add measures into their new modules that somehow balance out undesired model behavior. These measures can also be conceived as kluges.

Why should we see these examples as typical instances and not as exceptions? Because they arise from practical circumstances of developing software, which is a core part of simulation modeling. Software engineering is a field that was envisioned as the "professional" answer to the increasing complexity of software. And I frankly admit that there are well-articulated concepts that would in principle ensure software is clearly written, aptly modularized, well maintained, and superbly documented. However, the problem is that science *in principle* is different from science *in practice*.

In practice, there are strong and constant forces that drive software development into resorting to kluges. Economic considerations are always a reason, be it on the personal scale of research time, be it on the grand scale of assigning teams of developers to certain tasks. Usually, software is developed "on the move", i.e. those who write it have to keep up with changing requirements and a narrow timeline, in science as well as industry. Of course, in the ideal case the implementation is tightly modularized. A virtue of modularity is that it is much quicker incorporating "foreign" modules than developing them from scratch.

If these modules have some deficiencies, however, the developers will usually not start a fundamental analysis of how unexpected deviations occurred, but rather spend their energy for

adapting the interfaces so that the joint model will work as anticipated in the given circumstances. In common language: repair, rather than replace. Examples reach from integrating a module of atmospheric chemistry into an existing general circulation model up to implementing the new version of the operating system of your computer. Working with complex computational and simulation models seems to require a certain division of labor and this division, in turn, thrives on software traveling easily. At the same time, this will provoke kluges on the side of those that try to connect software modules.

Kluges thus arise from unprincipled reasons: throw-away code, made for the moment, is not replaced later, but becomes forgotten, buried in more code, and eventually fixed. This will lead to a cascade of kluges. Once there, they prompt more kluges, tending to become layered and entrenched.[10]

Foote and Yoder, prominent leaders in the field of software development, give an ironic and funny account of how attempts to maintain a rationally designed software architecture constantly fail in practice.

"While much attention has been focused on high-level software architectural patterns, what is, in effect, the de-facto standard software architecture is seldom discussed. This paper examines this most frequently deployed of software architectures: the BIG BALL OF MUD. A big ball of mud is a casually, even haphazardly, structured system. Its organization, if one can call it that, is dictated more by expediency than design. Yet, its enduring popularity cannot merely be indicative of a general disregard for architecture. (…) 2. Reason for degeneration: ongoing evolutionary pressure, piecemeal growth: Even systems with well-defined architectures are prone to structural erosion. The relentless onslaught of changing requirements that any successful system attracts can gradually undermine its structure. Systems that were once tidy become overgrown as piecemeal growth gradually allows elements of the system to sprawl in an uncontrolled fashion." (Foote and Yoder 1999, ch. 29)

I would like to repeat the statement from above that there is no necessity in the corruption of modularity and rational architecture. Again, this is a question of science in practice vs. science in principle. "A sustained commitment to refactoring can keep a system from subsiding into a big

[10] Wimsatt (2007) writes about "generative entrenchment" when speaking about the analogy between software development and biological evolution, see also Lenhard and Winsberg (2010).

ball of mud," Foote and Yoder concede. There are even directions in software engineering that try to counteract the degradation into Foote's and Yoder's big ball of mud. The movement of "clean code", for instance, is directed against what Foote and Yoder describe. Robert Martin, the pioneer of this school, proposes to keep code clean in the sense of not letting the first kluge slip in. And surely there is no principled reason why one should not be able to avoid this. However, even Martin accepts the diagnosis of current practice.

Similarly, Richard Gabriel (1996), another guru of software engineering, makes the analogy to housing architecture and Alexander's concept of "habitability", which intends to integrate modularity and piecemeal growth into one "organic order". Anyway, when describing the starting point, he more or less duplicates what we heard above from Foote and Yoder.

Finally, I want to point out that the matter of kluging is related to what is discussed in philosophy of science under the heading of opacity (like in Humphreys 2009). Highly kluged software becomes opaque. One can hardly disentangle the various reasons that led to particular pieces of code, because kluges are sensible only in the particular context at the time. In this important sense, simulation models are historical objects. They carry around – and depend on – their history of modifications. There are interesting analogies with biological evolution that have become a topic when complex systems had become a major issue in discussion computer use. Winograd and Flores, for instance, come to a conclusion that also holds in our context here: "each detail may be the result of an evolved compromise between many conflicting demands. At times, the only explanation for the system's current form may be the appeal to this history of modification." (1991, 94)[11]

Thus, the brief look into the somewhat elusive field of software development has shown us that two conditions foster kluging. First, the exchange of software parts that is more or less motivated by flexibility and economic requirements. This thrives on networked infrastructure. Second, iterations and modifications are easy and cheap. Due to the unprincipled nature of kluges, their construction requires repeated testing whether they actually work in the factual circumstances. Kluges hence fit to the exploratory and iterative mode of modeling that characterizes

[11] Interestingly, Jacob (1994) gives a very similar account of biological evolution when he writes that simpler objects are more dependent on (physical) constraints than on history, while history plays the greater part when complexity increases.

simulations. Furthermore, layered kluges solidify themselves. They make code hard or impossible to understand; modifying pieces that are individually hard to understand will normally lead to a new layer of kluges – and so on. Thus, kluging makes modularity erode and this is the second argument why simulation modeling systematically undermines modularity.

5. The limits of validation

What does the erosion of modularity mean for the validation of computer simulations? We have seen that the power and scope of simulation is built on the tendency toward holism. But holism and the erosion of modularity are two sides of the same coin. The key point regarding methodology is that holism is driven by the very procedure that makes simulation so widely applicable! It is through adjustable parameters that simulation models can be applied to systems beyond the control of theory (alone). It is through this very strategy that modularity erodes.

One ramification of utmost importance is about the concept of validation. In the context of simulation models the community speaks of verification and validation, or "V&V". Both are related, but the unanimous advice in the literature is to keep them separate. While verification checks the model internally, i.e. whether the software indeed captures what it is supposed to, validation checks whether the model adequately represents the target system. A standard definition states that "verification [is] the process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model." While validation is defined as "the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model." (Oberkampf and Trucano 2000, 3) Though there is some leeway of defining V&V, you get the gist of it from the saying: verification checks whether the model is right[12], while validation checks whether we have the right model.

Due to the increasing usage and growing complexity of simulations, the issue of V&V is itself a growing field in simulation literature. One example is the voluminous monograph by Oberkampf and Roy (2010) that meticulously defines and discusses the various steps to be included in V&V procedures. A first move in this analysis is to separate model form from model parameters. Each

[12] This sloppy saying should not obscure that the process of verification comprises a package of demanding tasks.

parameter then belongs to a particular type of parameter that determines which specific steps in V&V are required. Oberkampf gives the following list of model parameter types:

> - measurable properties of the system or the surroundings,
> - physical modeling parameters,
> - ad hoc model parameters,
> - numerical algorithm parameters,
> - decision parameters,
> - uncertainty modeling parameters." (Oberkampf and Roy 2010, section 13.5.1, p.623)

My point is that the adjustable parameters we discussed are of a type that is evading the V&V fencing. These parameters cannot be kept separate from the model form, since the form alone does not capture representational (nor behavioral) adequacy. A cloud parameterization scheme makes sense only with parameter values already assigned and the same holds for a many-parameter density functional. Before the process of adjustment, the mere form of the functional does not offer anything to be called adequate or inadequate. In simulation models, as we have seen, (predictive) success and adaptation are entangled.

The separation of verification and validation thus cannot be fully maintained in practice. It is not possible to first verify that a simulation model is 'right' before tackling the 'external' question whether it is the right model. Performance tests hence become the main handle for confirmation. This is a version of confirmation holism that points toward the limits of analysis. This does not lead to a complete conceptual breakdown of verification and validation. Rather, holism comes in degrees[13] and is a pernicious tendency that undermines the verification-validation divide.[14]

Finally, we come back to the analogy, or rather dis-analogy between computer and clockwork. In an important sense, computers are not amplifiers, i.e. they are not analogous to gigantic clockworks. They do not (simply) amplify the force of mathematical modeling that has got stuck

[13] I thank Rob Muir for pointing this out to me.

[14] My conclusion about the inseparability of verification and validation is in good agreement with Winsberg's more specialized claim in (2010) where he argues about model versions that evolve due to changing parameterizations, which has been criticized by Morrison (2015). As far as I can see, her arguments do not apply to the case made in this paper, which rests on a tendency toward holism, rather than a complete conceptual breakdown.

in too demanding operations. Rather, computer simulation is profoundly *changing* the setting of how mathematics is used.

In the present paper I questioned the rational picture of design. Also Brooks did this when he observed that Pahl and Beitz had to include more and more steps to somehow capture an unwilling and complex practice of design, or when he refers to Donald Schön who criticized a one-sided "technical rationality" that underlies the Rational Model (Brooks 2010, chapter 2). However, my criticism works, if you want, from 'within'. It is the very methodology of simulation modeling, and how it works in practice, that challenges the rational picture by making modularity erode.

The challenge to the rational picture has quite fundamental ramification because this picture influenced so many ways we conceptualize our world. I will spare the philosophical discussion of how simulation modeling is challenging our concept of mathematization and with it our picture of scientific rationality for another paper. Just let me mention the philosophy of mind as one example. How we are inclined to think about mind today is deeply influenced by the computer and by our concept of mathematical modeling. Jerry Fodor has defended a most influential thesis that mind is composed of information-processing devices that operate largely separately (Fodor 1983). Consequently, re-thinking how computer models are related to modularity invites to re-thinking the computational theory of the mind.

I would like to thank …

References

Agre, Philip E., Hierarchy and History in Simon's "Architecture of Complexity", journal of the learning sciences, 3, 2003, 413-426.

Brooks, Frederic P, The Design of Design. Boston, MA: Addison-Wesley, 2010.

Clark, Andy, The Kludge in the Machine, in: Mind and Language 2(4), 1987, 277-300.

Fillion, Nicolas, 2017, The Vindication of Computer Simulations, in Lenhard, J., and Carrier, M. (eds.), *Mathematics as a Tool*, Boston Studies in History and Philosophy of Science, forthcoming.

Fodor, Jerry: *The Modularity of Mind*, 1983, MIT Press, Cambridge, MA.

Foote, Brian und Joseph Yoder, *Pattern Languages of Program Design 4* (= *Software Patterns*. 4). Addison Wesley, 1999.

Frigg, Roman and Julian Reiss, The Philosophy of Simulation. Hot New Issues or Same Old Stew?, in: *Synthese*, 169(3), 593-613, 2009.

Gabriel, Richard P.: *Patterns of Software. Tales From the Software Community*, New York and Oxford: Oxford University Press, 1996.

Gramelsberger, Gabriele und Johann Feichter (eds.): *Climate Change and Policy. The Calculability of Climate Change and the Challenge of Uncertainty*, Heidelberg: Springer 2011.

Hasse, Hans, and Lenhard, J. (2017), On the Role of Adjustable Parameters, in Lenhard, J., and Carrier, M. (eds.), *Mathematics as a Tool*, Boston Studies in History and Philosophy of Science, forthcoming.

Humphreys, Paul, The Philosophical Novelty of Computer Simulation Methods, *Synthese*, 169 (3):615 - 626 (2009).

Jacob, Francois, *The Possible and the Actual*, Seattle: University of Washington Press, 1994.

Lenhard, Johannes, *Disciplines, Models, and Computers: The Path To Computational Quantum Chemistry*, Studies in History and Philosophy of Science Part A, 48 (2014), 89-96.

Lenhard, Johannes and Eric Winsberg, *Holism, Entrenchment, and the Future of Climate Model Pluralism*, in: Studies in History and Philosophy of Modern Physics, 41, 2010, 253-262.

Mauritsen, Thorsten, Bjorn Stevens, Erich Roeckner, Traute Crueger, Monika Esch, Marco Giorgetta, Helmuth Haak, Johann Jungclaus, Daniel Klocke, Daniela Matei, Uwe Mikolajewicz, Dirk Notz, Robert Pincus, Hauke Schmidt, and Lorenzo Tomassini, Tuning the climate of a global model, *Journal of Advances in Modeling Earth Systems*, 4, 2012.

Morrison, Margaret, *Reconstructing Reality. Models, Mathematics, and Simulations*. New York: Oxford University Press, 2015.

Oberkampf, William L., and Roy, Christopher J., *Verification and Validation in Scientific Computing*. Cambridge, MA: Cambridge University Press, 2010.

Oberkampf, William L. and Trucano, T.G., *Validation Methodology in Computational Fluid Dynamics*, American Institute for Aeronautics and Astronautics, 2000 – 2549, 2000.

Pahl, G. and Beitz, W. 1984. *Engineering Design: A Systematic Approach*. Revised editions in 1996, 2007. Berlin: Springer.

Parker, Wendy, Values and Uncertainties in Climate Prediction, revisited, *Studies in History and Philosophy of Science* 2013.

Perdew, J. P., Ruzsinsky, A., Tao, J., Staroverov, V., Scuseria, G., & Csonka, G. (2005). Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *The Journal of Chemical Physics*, 123.

Royce, Wynston, Managing the Development of Large software Systems. *Proceedings of IEEE WESCON* 26 (August), 1970, 1–9.

Simon, Herbert A., *The Sciences of the Artificial*, Cambridge, MA: The MIT Press, 1969.

Smith, Leonard A., What Might We learn From Climate Forecasts?, in: *Proceeedings of the National Academy of Sciences USA*, 4(99), 2002, 2487-2492.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.), *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel*

on Climate Change, 2007. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Stainforth, D.A., Downing, T.E., Washington, R. and New, M. (2007) Issues in the interpretation of climate model ensembles to inform decisions, *Philosophical Transactions of the Royal Society*, Volume 365, Number 1857, 2145-2161.

Wimsatt, William C., *Re-Engineering Philosophy for Limited Beings. Piecewise approximations to reality*, Cambridge, MA and London, England: Harvard University Press, 2007.

Winograd, Terry und F. Flores, *Understanding Computers and Cognition*, Reading, MA: Addison-Wesley, ^{5}1991.

Winsberg, Eric, *Science in the Age of Computer Simulation*, Chicago, Ill.: University of Chicago Press, 2010.

Accuracy, conditionalization, and probabilism

Peter J. Lewis, University of Miami
Don Fallis, University of Arizona

March 3, 2016

Abstract

Accuracy-based arguments for conditionalization and probabilism appear to have a significant advantage over their Dutch Book rivals. They rely only on the plausible epistemic norm that one should try to decrease the inaccuracy of one's beliefs. Furthermore, it seems that conditionalization and probabilism follow from a wide range of measures of inaccuracy. However, we argue that among the measures in the literature, there are some from which one can prove conditionalization, others from which one can prove probabilism, and none from which one can prove both. Hence at present, the accuracy-based approach cannot underwrite both conditionalization and probabilism.

A central concern of epistemology is uncovering the rational constraints on an agent's credences, both at a time and over time. At a time, it is typically maintained that an agent's credences should conform to the probability axioms, and over time, it is often maintained that an agent's credences should conform to conditionalization. How could such norms be justified? The traditional approach is to show that if your credences violate these norms, then there is a set of bets, each of which you consider fair, but which collectively are such that if you accept them all you will lose money whatever happens. Since you do not want to be a "money pump", you should adopt coherent credences. However, this *Dutch book* strategy rests on controversial assumptions concerning prudential rationality and its connection to epistemic rationality.

The prudential elements may not be essential to the Dutch book approach (Vineberg 2012). But even so, it would be better to be able to derive probabilism and conditionalization from a clearly epistemic basic norm. A more

recent approach seeks to do precisely that: to derive probabilism and conditionalization from the intuitive epistemic norm that you should endeavor to make your credences as accurate—as close to the truth—as possible. Drawing on the work of Joyce (1998; 2009), Greaves and Wallace (2006) and Predd et al. (2009), Pettigrew (2013) argues that the accuracy-based approach vindicates both probabilism and conditionalization. We argue that this conclusion is too strong: at present, the accuracy-based approach can vindicate *either* conditionalization *or* probabilism, but not both.

Our argument turns on the features of various proposed measures of accuracy. The accuracy-based approach is predicated on the assumption that the accuracy of your credences can be measured. Pettigrew (2013, 905) argues that it is a strength of the accuracy-based approach that conditionalization and probabilism follow from a wide range of measures, so that it doesn't matter which measure is used to assess the accuracy of an agent's credences. Our counter-argument is that it does matter: of the known measures, some vindicate conditionalization, and some vindicate probabilism, but there is no known measure of inaccuracy from which both conditionalization and probabilism can be derived.

1 Accuracy and conditionalization

First, let us briefly run through the argument via which conditionalization and probabilism are claimed to follow from considerations of accuracy, starting with conditionalization. Suppose you have credences $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ in propositions $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, where the propositions form a partition, i.e. they are exhaustive and mutually exclusive, so that exactly one of them is true. The accuracy approach takes it that your primary epistemic goal is having credences that are as accurate as possible, where complete accuracy is a credence of 1 in the true proposition and a credence of 0 in each of the false propositions. The closer your credences are to complete accuracy, the better.

For this epistemic goal to make sense, we need a measure of closeness. In what follows we will discuss several such measures, expressed as measures of *inaccuracy*: the larger the measure, the further your credences are from the truth. Hence your goal is to minimize the value of this inaccuracy measure. By far the dominant measure in the literature is the quadratic rule or Brier rule, which takes the square of the difference between your credence in each

proposition and its truth value, and sums the results. So for a partition, if $I_i(\mathbf{b})$ is the inaccuracy of credences $\mathbf{b}$ when proposition X_i is true, then the Brier rule can be expressed as follows:[1]

Simple Brier rule: $I_i(\mathbf{b}) = (1 - b_i)^2 + \sum_{j \neq i} b_j^2$.

The Brier rule has been defended by epistemologists (Joyce 2009, 290; Leitgeb and Pettigrew 2010, 219), and is frequently cited as the prime example of an inaccuracy measure (Greaves and Wallace 2006, 627; Pettigrew 2013, 899).

Suppose you obtain evidence E that is consistent with some but not all of the propositions $\mathbf{X}$. How should you distribute your credence over the remaining propositions? If your goal is to minimize your inaccuracy, presumably the best you can do is to minimize your *expected* inaccuracy given your prior credences $\mathbf{b}$. So suppose that after you learn E, you shift your credence in proposition X_i from b_i to x. If X_i is true, the contribution of this new credence to your overall inaccuracy is $(1 - x)^2$, and if X_i is false, the contribution is x^2. Given your prior credences $\mathbf{b}$, you judge that the chance that X_i is true is b_i, and the chance that X_i is false is $\sum_{E-i} b_j$, where the notation $E - i$ indicates that the sum is over all propositions consistent with E except X_i. That is, the total contribution C of this new credence to your expected inaccuracy is given by:

$$C = (1 - x)^2 \, b_i + x^2 \sum_{E-i} b_j.$$

Your goal is to minimize C. So consider where $dC/dx = 0$:

$$\begin{aligned} \frac{dC}{dx} &= -2\,(1 - x)\, b_i + 2x \sum_{E-i} b_j \\ &= -2b_i + 2x \sum_{E} b_j, \end{aligned}$$

where the sum in the last line is now over all propositions consistent with E. This expression is zero when

$$x = \frac{b_i}{\sum_E b_j}.$$

[1] We call the version of the Brier rule applicable to a partition the *simple* Brier rule only for ease of reference (and similarly for the simple log rule and simple spherical rule to be introduced later).

But note that this value for x is just your prior credence in X_i conditional on E:

$$c(X_i|E) = \frac{c(X_i \wedge E)}{c(E)} = \frac{b_i}{\sum_E b_j}.$$

That is, conditionalizing on E minimizes your expected inaccuracy.[2] So if your epistemic goal is to minimize inaccuracy, you should conditionalize on new evidence.

Greaves and Wallace (2006) generalize this proof to cover measures of inaccuracy other than the Brier rule. In particular, they show that conditionalization minimizes expected inaccuracy for any measure of inaccuracy $I_i(\mathbf{b})$ satisfying *strict propriety*:

Strict propriety: For any distinct probabilistic credences $\mathbf{b}$ and $\mathbf{b'}$, $\sum_i b_i I_i(\mathbf{b}) < \sum_i b_i I_i(\mathbf{b'})$.

Strict propriety says that the expected inaccuracy of your current credences $\mathbf{b}$ is lower than the expected inaccuracy of any alternative credences $\mathbf{b'}$ you might adopt, where the expectation is calculated according to your current credences. If it fails, then the injunction to minimize inaccuracy makes your beliefs pathologically unstable: you can lower your expected inaccuracy by shifting your credences, even in the absence of new evidence. Hence strict propriety serves as a reasonable constraint on measures of inaccuracy. The Brier rule is strictly proper, as are several other proposed inaccuracy measures to be discussed below.

Greaves and Wallace begin by introducing some terminology. They say that a set of credences $\mathbf{b}$ *recommends* a set of credences $\mathbf{b'}$ iff the expected inaccuracy of $\mathbf{b'}$ is at least as low as the expected inaccuracy of $\mathbf{b}$, where the expectation is calculated using credences $\mathbf{b}$:

Recommendation: $\mathbf{b}$ recommends $\mathbf{b'}$ iff $\sum_i b_i I_i(\mathbf{b}) \geq \sum_i b_i I_i(\mathbf{b'})$

Note that if the inaccuracy measure $I_i(\mathbf{b})$ satisfies strict propriety, then $\mathbf{b}$ only recommends itself.

They further define *quasi-conditionalization* as a belief updating rule that stipulates that your credences on learning E should be some set *recommended* by your prior credences conditional on E. They then prove

[2] This proof is a simplified version of the one in Leitgeb and Pettigrew (2010).

that quasi-conditionalization is always optimal: whatever measure of inaccuracy you choose, strictly proper or not, the expected inaccuracy of quasi-conditionalizing is at least as low as the expected inaccuracy of any other updating rule. Then if your measure of inaccuracy is strictly proper, conditionalization itself is optimal, since for strictly proper measures, credences only recommend themselves. In fact, since the inequality in strict propriety is *strict*, conditionalization is strictly better than any other updating rule: it uniquely minimizes expected inaccuracy. As Pettigrew (2013, 905) notes, this is a strong result: any inaccuracy measure satisfying strict propriety can be used to vindicate conditionalization, and strict propriety is a constraint we would expect any reasonable inaccuracy measure to obey anyway.

2 Accuracy and probabilism

Now let us turn to the arguments that your credences at a time should obey the probability axioms. So far, we have been assuming that the propositions we are interested in form a partition. But the probability axioms include constraints on your credences in disjunctions, and to model such constraint we need to allow that more than one of the propositions you are considering can be true. To that end, suppose that you have credences $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ in propositions $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, where now the set of propositions forms a Boolean algebra, i.e. it is closed under negation and disjunction. So now we can no longer model a possible world simply as an index (picking out the unique true proposition); instead, we need to label each proposition separately as either true or false. That is, a possible world is specified by $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots \omega_n)$, where $\omega_i = 1$ when X_i is true and $\omega_i = 0$ when X_i is false. In this context, the Brier rule can be rewritten as follows:

Symmetric Brier rule: $I(\boldsymbol{\omega}, \mathbf{b}) = \sum_i (b_i - \omega_i)^2$.

As before, the inaccuracy of your beliefs according to the Brier rule is given by the sum of the squares of the distance of each belief from the relevant truth value. That is, the Brier rule is *symmetric*, in the sense that distance from the truth for a true proposition plays the same role as distance from falsity plays for a false proposition. This property will be important later.

The general strategy for defending probabilism based on accuracy goes as follows. Suppose that your current credences are incoherent—that is, they

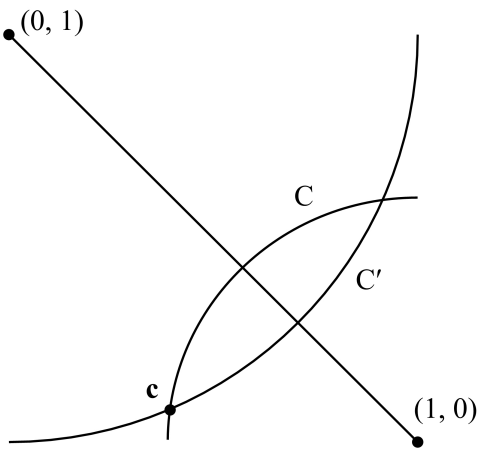


Figure 1: De Finetti's construction for a two-element partition (Joyce 1998, 582).

violate the probability axioms. Then one can appeal to a measure of inaccuracy to show that there are coherent credences that *dominate* your current credences—that are more accurate than your current credences whatever the truth values of the propositions concerned. If your goal is to minimize inaccuracy, this gives you a clear reason to avoid incoherent credences: there are always coherent credences that are more accurate, whatever the world is like.

De Finetti (1974, 87) constructs a dominance argument of this kind based on the Brier rule.[3] For illustration, consider the simple case of a proposition and its negation: that is, the propositions under consideration are just $(X, \neg X)$. In this case the space of possible credences forms a plane, as shown in figure 1: your credence in X is the horizontal coordinate, and your credence in $\neg X$ is the vertical coordinate. The two possible worlds are represented by the points $(1, 0)$ and $(0, 1)$, and your credences obey the probability axioms if and only if they lie on the straight line that connects these two points, since along this line your credences in X and $\neg X$ sum to 1.

Suppose that your credences are incoherent: they are represented by a point $\mathbf{c} = (c_1, c_2)$ that lies *off* this diagonal. And suppose first that the

[3] As Joyce (1998, 580) notes, de Finetti sets up this argument in terms of bets. However, as Pettigrew (2013, 901) points out, it can be redescribed as an accuracy-based argument.

actual world is represented by the bottom-right corner $(1, 0)$—i.e. X is true and $\neg X$ is false. Then the inaccuracy of your credences according to the Brier rule is $I(\boldsymbol{\omega}, \mathbf{c}) = (1 - c_1)^2 + (c_2)^2$. Note that this is just the square of the Euclidean distance between (c_1, c_2) and $(1, 0)$. That is, every point on the circle segment C has the same inaccuracy as $\mathbf{c}$, and every point between C and $(1, 0)$ has a lower inaccuracy. Now suppose instead that the actual world is represented by the top-left corner $(0, 1)$—i.e. X is false and $\neg X$ is true. Then the inaccuracy of your credences is $I(\boldsymbol{\omega}, \mathbf{c}) = (c_1)^2 + (1 - c_2)^2$—the square of the Euclidean distance between (c_1, c_2) and $(0, 1)$. That is, every point on the circle segment C' has the same inaccuracy as $\mathbf{c}$, and every point between C and $(0, 1)$ has a lower inaccuracy.

Consider the area enclosed by the circle segments C and C'. The credences represented by the points in this area have a lower inaccuracy than $\mathbf{c}$ if X is true and $\neg X$ false, and a lower inaccuracy than $\mathbf{c}$ if X is false and $\neg X$ true. That is, they have a lower inaccuracy whatever the world is like. And this area includes part of the diagonal that represents coherent credences. So for any incoherent set of credences, there is a coherent set that is less inaccurate whatever the world is like. In this simple case, accuracy gives you a motive to adopt coherent credences.

In the general case, the space of possible credences is n-dimensional, where there are n propositions in the Boolean algebra. Each possible assignment of truth values to the n propositions is represented by a point in this space, and the set of coherent credences consists of these points plus the points on the straight lines that connect them, the points on the straight lines that connect those latter points, and so on. This set is called the *convex hull* V^+ of the possible truth value assignments V. Via a generalization of the construction of figure 1, de Finetti shows that if your credences are represented by a point that lies outside V^+, then there are points in V^+ that are more accurate (according to the Brier rule) whichever point in the space represents the actual truth values of the propositions. Hence if you have incoherent credences, there are always coherent credences with a lower inaccuracy as measured by the Brier rule.

Predd et al. (2009) generalize this proof strategy to cover a wider range of inaccuracy measures. Their proof relies on two assumptions. The first is additivity:

Additivity: $I(\boldsymbol{\omega}, \mathbf{b})$ can be expressed as $\sum_i s(\omega_i, b_i)$, where s is a continuous function of your credence in proposition X_i and its truth value.

Additivity states that the inaccuracy of your beliefs in a set of propositions is just the sum of your inaccuracies in the propositions taken individually—that is, $s(\omega_i, b_i)$ is the inaccuracy of your belief in proposition X_i, and $I(\boldsymbol{\omega}, \mathbf{b})$ is just the sum of these inaccuracies for all the propositions you are considering. Note that it also contains the requirement that the inaccuracy measure should be continuous. The Brier rule is obviously additive, since it is expressed as a sum over propositions.

The second assumption is a version of strict propriety. For an additive inaccuracy measure, strict propriety can be expressed in terms of your inaccuracy function for a single proposition $s(b_i, \omega_i)$ as follows:

Strict propriety (for an additive measure): $b_i s(x, 1) + (1 - b_i)s(x, 0)$ is uniquely minimized at $x = b_i$.

Predd et al. (2009) prove that any additive, strictly proper inaccuracy measure entails probabilism. De Finetti's construction appeals to the natural distance measure implicit in the Brier rule—the Euclidean distance between two points in the space of your possible credences. But in the current case we have no explicit measure of inaccuracy, so Predd et al. appeal to a generalized "distance" measure[4] called the Bregman divergence, defined for a strictly convex function $\Phi(\mathbf{x})$ as $d_\Phi(\mathbf{y}, \mathbf{x}) = \Phi(\mathbf{y}) - \Phi(\mathbf{x}) - \nabla\Phi(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x})$. They show that if the inaccuracy measure $s(b_i, \omega_i)$ for a single proposition X_i is strictly proper, then the function $\varphi(b_i) = -b_i s(b_i, 1) - (1 - b_i)s(b_i, 0)$ is strictly convex. In terms of this function, Predd et al. show that for any additive, strictly proper inaccuracy measure, $I(\boldsymbol{\omega}, \mathbf{b}) = d_\Phi(\boldsymbol{\omega}, \mathbf{b})$, where $\Phi(\boldsymbol{\omega}) = \sum_i \varphi(\omega_i)$ and $\Phi(\mathbf{b}) = \sum_i \varphi(b_i)$.

The set of coherent credences forms a closed, convex subspace V^+ of the space of all possible credences. It is a fact from the theory of Bregman divergences that for any point $\mathbf{c}$ outside V^+, there is a unique point $\mathbf{c}^*$ in V^+ such that $d_\Phi(\mathbf{c}^*, \mathbf{c}) \leq d_\Phi(\mathbf{y}, \mathbf{c})$ for all $\mathbf{y}$ in V^+. That is, $\mathbf{c}^*$ is the unique closest point in V^+ to $\mathbf{c}$, using the Bregman divergence as a distance measure. It is a further fact that $d_\Phi(\mathbf{y}, \mathbf{c}^*) \leq d_\Phi(\mathbf{y}, \mathbf{c}) - d_\Phi(\mathbf{c}^*, \mathbf{c})$ for all $\mathbf{y}$ in V^+ and $\mathbf{c}$ outside V^+. Note in particular that V^+ contains every possible world $\boldsymbol{\omega}$, since a consistent truth value assignment is also a coherent set of credences. So setting $\mathbf{y} = \boldsymbol{\omega}$, we have $d_\Phi(\boldsymbol{\omega}, \mathbf{c}^*) \leq d_\Phi(\boldsymbol{\omega}, \mathbf{c}) - d_\Phi(\mathbf{c}^*, \mathbf{c})$. Since d_Φ is a positive-valued function, $d_\Phi(\mathbf{c}^*, \mathbf{c}) > 0$, so $d_\Phi(\boldsymbol{\omega}, \mathbf{c}^*) < d_\Phi(\boldsymbol{\omega}, \mathbf{c})$, and hence

[4] The reason for the scare quotes is that the Bregman divergence is not symmetric, and distance measures are typically symmetric.

$I(\boldsymbol{\omega}, \mathbf{c}^*) < I(\boldsymbol{\omega}, \mathbf{c})$. That is, for any incoherent set of credences $\mathbf{c}$, there is a coherent set $\mathbf{c}^*$ that is less inaccurate than $\mathbf{c}$ in every possible world.

As Pettigrew (2013, 905) notes, this is a strong result: any inaccuracy measure satisfying strict propriety and additivity can be used to vindicate probabilism, and while additivity is perhaps not forced on us in the way that strict propriety is, it is certainly intuitive. As we shall see, there are several available measures satisfying additivity and strict propriety, so it initially looks like the accuracy-based program can justify both probabilism and conditionalization based on minimal premises. Our purpose in this paper is to argue that matters are not so straightforward.

3 Measures of inaccuracy

Let us return to the argument for conditionalization. This argument restricts inaccuracy measures to those that are strictly proper. Note that strict propriety is only a condition on *expected* inaccuracy. But expected inaccuracy is calculated on the basis of the *actual* inaccuracy that the measure in question ascribes to credences, and presumably there are a number of constraints any such measure must obey if it is to genuinely measure epistemic inaccuracy rather than something else. For example, if one of your credences shifts towards the truth, while your other credences stay the same, then clearly your actual inaccuracy should decrease. We wish to focus on one such constraint.

The constraint can be motivated by thinking about *elimination cases*. Suppose you are considering a set of mutually exclusive and exhaustive propositions, and suppose that your credences are coherent and that you conditionalize on evidence. You acquire some evidence that eliminates one false proposition—your credence in it becomes zero—but is uninformative regarding the other hypotheses—your credences in them remain in the same proportions. How does this affect the accuracy of your credences?

It seems obvious that your beliefs have become more accurate. If you believe that Tom, Dick or Harry might be the murderer (when in fact Tom did it), and you eliminate Harry while learning nothing about Tom or Dick, then you have made epistemic progress towards the truth, or at least away from falsity. It is true that your credence in the false proposition "Dick did it" goes up, but only by the same proportion that your credence in the true proposition "Tom did it" goes up.

Unfortunately, the simple Brier rule does not always concur. Let X_1 be

"Tom did it", X_2 be "Dick did it", and X_3 be "Harry did it", where unknown to you X_1 is true. Suppose that your initial credences in (X_1, X_2, X_3) are $\mathbf{b} = (1/7, 3/7, 3/7)$. Then according to the simple Brier rule, your initial inaccuracy is $54/49 = 1.10$. Now suppose you acquire some evidence that eliminates X_3, but is uninformative regarding X_1 and X_2. That is, your credence in X_3 becomes 0 and your credences in X_1 and X_2 stay in the same proportions, so that your final credences are $\mathbf{b}^* = (1/4, 3/4, 0)$. Then according to the simple Brier rule, your final inaccuracy is $18/16 = 1.13$. That is, the Brier rule erroneously says that the inaccuracy of your beliefs has gone up.

For a measure to genuinely measure the actual inaccuracy of your beliefs, it should not be susceptible to counterexamples of this kind; it should count elimination cases as epistemically positive. That is, measures of inaccuracy should obey the following principle:

M: For coherent credences over a partition, if $\mathbf{b}$ assigns a zero credence to some false proposition to which $\mathbf{b}'$ assigns a non-zero credence, and credences in the remaining propositions stay in in the same ratios, then $\mathbf{b}$ is epistemically better than $\mathbf{b}'$.

The simple Brier rule, as the example shows, violates M, and hence does not plausibly measure the actual inaccuracy of your beliefs.[5]

Fortunately, though, there are alternative inaccuracy measures for partitions we can appeal to. The two most frequently mentioned are the simple log rule and the simple spherical rule:

Simple log rule: $I_i(\mathbf{b}) = -\ln b_i$

Simple spherical rule: $I_i(\mathbf{b}) = 1 - b_i/\sqrt{\sum_j b_j^2}$.

As before, $I_i(\mathbf{b})$ is the inaccuracy of credences $\mathbf{b}$ when proposition X_i is true. Both of these measures satisfy M, and hence are not susceptible to elimination counterexamples.[6] Hence each can plausibly be claimed to measure epistemic inaccuracy. Furthermore, each is strictly proper, and so each can be used to

[5] One might reasonably think that acceptable measures of accuracy should obey a stronger principle than M; see (*reference removed*).

[6] This is trivial for the log rule, and easily proven for the spherical rule. See (*reference removed*).

underwrite conditionalization via the above argument strategy. So there are some inaccuracy measures that vindicate conditionalization, but not all strictly proper measures do so. In particular, the simple Brier rule cannot be used to vindicate conditionalization.

But what about probabilism? The simple log rule and simple spherical rule are not applicable to a Boolean algebra, and so cannot be used to prove probabilism as they stand. Perhaps the most straightforward way to generalize them is simply to sum the contribution given by the simple rule for each true proposition in the Boolean algebra, while ignoring the false propositions in the algebra:

Asymmetric log rule: $I(\mathbf{b}, \boldsymbol{\omega}) = \sum_i F(\omega_i, b_i)$, where $F(0, b_i) = 0$ and $F(1, b_i) = -\ln b_i$.

Asymmetric spherical rule: $I(\mathbf{b}, \boldsymbol{\omega}) = \sum_i F(\omega_i, b_i)$, where $F(0, b_i) = 0$ and $F(1, b_i) = 1 - b_i / \sqrt{\sum_j b_j^2}$.

Both these rules are asymmetric, in the sense that inaccuracy is calculated differently for true and false propositions. These rules satisfy principle M: for coherent credences, if your credence in a false proposition goes down and your remaining credences stay in the same ratios, then your credence in each true proposition goes up, and so your inaccuracy according to the relevant asymmetric rules goes down. Hence the asymmetric log and spherical rules are immune from elimination counterexamples.

But these rules do not satisfy the combination of additivity and strict propriety required for the proof of probabilism. The asymmetric spherical rule is not additive: $F(1, b_i)$ is not a function of b_i alone. The asymmetric log rule is additive, but it is not strictly proper in the required sense: $F(1, b_i)$ is strictly proper, but $F(0, b_i)$ is not. Indeed, it is straightforward to show directly that these rules cannot be used as the basis of a dominance argument for probabilism. Consider, for example, a two element partition, and the incoherent credence assignment $(1, 1)$. The asymmetric log rule counts these incoherent credences as *perfectly* accurate (since the credence in the false proposition is ignored), so no coherent credences can dominate them. According to the asymmetric spherical rule, multiplying all credences by a constant has no effect on inaccuracy, so this assignment has the same inaccuracy as the coherent credence assignment $(1/2, 1/2)$. If coherent assignments cannot be dominated, then neither can the initial incoherent assignment.

But if coherent assignments *can* be dominated then the dominance proof of probabilism fails anyway.

So the asymmetric versions of the log rule and the spherical rule cannot be used to prove probabilism. But for a Boolean algebra, the log rule and the spherical rule are usually given a formulation that is symmetric between truth and falsity:

Symmetric log rule: $I(\boldsymbol{\omega}, \mathbf{b}) = \sum_i - \ln |(1 - \omega_i) - b_i|$

Symmetric spherical rule: $I(\boldsymbol{\omega}, \mathbf{b}) = \sum_i 1 - \dfrac{|(1 - \omega_i) - b_i|}{\sqrt{b_i^2 + (1 - b_i)^2}}$

(see e.g. Joyce 2009, 275). These measures are additive, and each term in the sum is individually strictly proper, so they can each be used to prove probabilism via the proof of Predd et al.

But unfortunately, in their symmetric forms all three rules—Brier, log and spherical—are subject to elimination counterexamples. For the Brier rule, the counterexample is the same as before, since the symmetric Brier rule reduces to the simple Brier rule when applied to a partition.[7] That is, consider a credence shift from $\mathbf{b} = (1/7, 3/7, 3/7)$ to $\mathbf{b}^* = (1/4, 3/4, 0)$ when X_1 is true. According to the symmetric Brier rule, your initial inaccuracy is 1.10, and your final inaccuracy is 1.13, so your inaccuracy goes up. And this example works equally well against the symmetric spherical rule: according to this rule, your initial inaccuracy is 1.24 and your final inaccuracy is 1.37, so your inaccuracy goes up. This particular counterexample does not work against the symmetric log rule, but a similar one does. Suppose your initial credences are $\mathbf{b} = (1/13, 6/13, 6/13)$, and your final credences are $\mathbf{b}^* = (1/7, 6/7, 0)$. Then according to the symmetric log rule your initial inaccuracy is 3.80, and your final inaccuracy is 3.89: your inaccuracy goes up. Hence the symmetric measures all violate principle M, and so none of them can be used to prove conditionalization.

[7] Strictly, applying these rules to a Boolean algebra requires including credences in the negations $\neg X_1$, $\neg X_2$ and $\neg X_3$, plus the tautology $X_1 \vee X_2 \vee X_3$ and the contradiction $\neg(X_1 \vee X_2 \vee X_3)$. But for coherent credences the inaccuracies of the tautology and the contradiction are zero, and for symmetric rules the inaccuracy of $\neg X_i$ is the same as that of X_i, so the inaccuracy calculated over the entire Boolean algebra is simply twice the inaccuracy over the partition (X_1, X_2, X_3).

4 The extent of the problem

Let us sum up. The simple Brier rule cannot be used to prove conditionalization, but the simple log and spherical rules can. The obvious generalizations of the simple log and spherical rules to a Boolean algebra—the asymmetric log and spherical rules—cannot be used to prove probabilism. The symmetric Brier, log and spherical rules can be used to prove probabilism, but none of them underwrites conditionalization. So we have found no measure that can be used to prove both conditionalization *and* probabilism.

Could there be such a measure? Perhaps, although it is worth noting that one can prove that *any* inaccuracy measure that satisfies additivity, strict propriety and a plausible symmetry principle is subject to elimination counterexamples. The symmetry principle is precisely the one discussed above—that the inaccuracy measure treats truth the same as falsity, in the sense that it is a function of the distance between each credence and its respective truth value. For an additive inaccuracy measure, the symmetry principle can be expressed in terms of the inaccuracy function for a single proposition $s(\omega_i, b_i)$ as follows:

Symmetry: $s(\omega_i, b_i) = s(|1 - \omega_i|, |1 - b_i|)$.

It is certainly highly plausible that this is part of what it means for s to measure your distance from the truth, and as discussed above, the typical Boolean algebra forms of the Brier rule, log rule and spherical rule all satisfy it.

Let us see how this symmetry principle, together with additivity and strict propriety, lead to elimination counterexamples. Consider a single proposition X_i in which your credence is $b_i = 1/2$. According to strict propriety, the quantity $(1/2)s(1, x) + (1/2)s(0, x)$ must be uniquely minimized at $x = 1/2$. In particular, the value of this expression for $x = 1/2$ must be lower than its value for $x = 1$:

$$(1/2)s(1, 1/2) + (1/2)s(0, 1/2) < (1/2)s(1, 1) + (1/2)s(0, 1),$$

and for $x = 0$:

$$(1/2)s(1, 1/2) + (1/2)s(0, 1/2) < (1/2)s(1, 0) + (1/2)s(0, 0).$$

Adding these:

$$s(1, 1/2) + s(0, 1/2) < (1/2)s(1, 1) + (1/2)s(0, 1) + (1/2)s(1, 0) + (1/2)s(0, 0).$$

But by symmetry, $s(1,1/2) = s(0,1/2)$, $s(1,1) = s(0,0)$ and $s(0,1) = s(1,0)$. Substituting:

$$2s(0,1/2) < s(0,1) + s(0,0).$$

Now consider your credences in three exhaustive and mutually exclusive propositions $\mathbf{X} = (X_1, X_2, X_3)$. Consider in particular the credence shift from $\mathbf{m} = (0, 1/2, 1/2)$ to $\mathbf{b} = (0, 1, 0)$ for truth values $\boldsymbol{\omega} = (1, 0, 0)$. By separability, $I(\boldsymbol{\omega}, \mathbf{m}) = s(1,0) + 2s(0,1/2)$, and $I(\boldsymbol{\omega}, \mathbf{b}) = s(1,0) + s(0,1) + s(0,0)$. So since $2s(0,1/2) < s(0,1) + s(0,0)$ it follows that $I(\boldsymbol{\omega}, \mathbf{m}) < I(\boldsymbol{\omega}, \mathbf{b})$: your inaccuracy goes up. But the shift from $\mathbf{m} = (0, 1/2, 1/2)$ to $\mathbf{b} = (0, 1, 0)$ is an elimination case: a false proposition is eliminated, and your credences in the remaining hypotheses stay in the same proportions. And lest one worry about the fact that your initial credence in the true proposition is zero, we can modify the example. Consider the credence assignments $\mathbf{m}' = (\delta/(2+\delta), 1/(2+\delta), 1/(2+\delta))$ and $\mathbf{b}' = (\delta/(1+\delta), 1/(1+\delta), 0)$. For small δ these are close to $\mathbf{m}$ and $\mathbf{b}$, and hence by the continuity clause of additivity, the inaccuracy of $\mathbf{m}'$ remains lower than that of $\mathbf{b}'$. Again, the transition from $\mathbf{m}'$ to $\mathbf{b}'$ is an elimination case, and now your credence in the true proposition is non-zero.

So elimination counterexamples afflict any inaccuracy measure that satisfies additivity, strict propriety and symmetry. That is, any symmetric measure that satisfies the assumptions of Predd et al.'s proof of probabilism violates principle M, and hence cannot be used to prove conditionalization. Symmetry is not a premise in the Predd argument, so it is possible that an asymmetric measure might allow the derivation of both probabilism and conditionalization. But the only plausible asymmetric measure in the literature is the log rule (Bernardo 1979), and we have seen that the asymmetric log rule does not vindicate probabilism.

5 Conclusion

Pettigrew notes that conditionalization and probabilism follow from a wide range of measures of inaccuracy, and the implication is that it doesn't much matter which measure you pick. But we think it does matter. There are measures that vindicate conditionalization, and there are measures that vindicate probabilism, but nobody has yet identified a measure that vindicates both. Hence the accuracy-based approach does not, as yet, give us the justification we might want for the constraints on our credences.

References

Bernardo, José M. (1979), "Expected information as expected utility", *Annals of Statistics* 7: 686-690.

de Finetti, Bruno (1974), *Theory of Probability*, vol. 1. New York: John Wiley and Sons.

Greaves, Hilary and David Wallace (2006), "Justifying conditionalization: conditionalization maximizes expected epistemic utility", *Mind* 115: 607–32.

Joyce, James M.. (1998), "A nonpragmatic vindication of probabilism", *Philosophy of Science*, 65: 575–603.

Joyce, James M. (2009), "Accuracy and coherence: prospects for an alethic epistemology of partial belief", in F. Huber and C. Schmidt-Petri (eds.), *Degrees of Belief*. Dordrecht: Springer: 263–97.

Leitgeb, Hannes, and Richard Pettigrew (2010), "An objective justification of Bayesianism I: measuring inaccuracy", *Philosophy of Science* 77: 201–35.

Pettigrew, Richard (2013), "Epistemic utility and norms for credence", *Philosophy Compass* 8: 897–908.

Predd, Joel B., Robert Seiringer, Elliott H. Lieb, Daniel N. Osherson, H. Vincent Poor, and Sanjeev R. Kulkarni (2009), "Probabilistic coherence and proper scoring rules", *IEEE Transactions on Information Theory* 55: 4786–4792.

Vineberg, Susan (2012), "Dutch book arguments", in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/spr2012/entries/dutch-book/

Can Typicality Arguments Dissolve Cosmology's Flatness Problem?

C.D. McCoy*

20 February 2016

Abstract

The flatness problem in cosmology draws attention to a surprising fine-tuning of the spatial geometry of our universe towards flatness. Several physicists, among them Hawking, Page, Coule, and Carroll, have argued against the probabilistic intuitions underlying such fine-tuning arguments in cosmology and instead propose that the canonical measure on the phase space of Friedman-Robertson-Walker spacetimes should be used to evaluate fine-tuning. They claim that flat spacetimes in this set are actually typical on this natural measure and that therefore the flatness problem is illusory. I argue that they misinterpret typicality in this phase space and, moreover, that no conclusion can be drawn at all about the flatness problem by using the canonical measure alone.

For several decades now cosmologists have maintained that the old standard model of cosmology, the highly successful hot big bang (HBB) model, suffers from various fine-tuning problems (Dicke and Peebles, 1979; Linde, 1984). They claim that the spacetimes on which the HBB model is based, the Friedman-Robertson-Walker (FRW) spacetimes, require seemingly "special" initial conditions, such that when they are evolved forward in time by the dynamical law of the general theory of relativity (GTR) they yield presently observed cosmological conditions. For example, the flatness problem depends on the existence of special initial conditions in the HBB model which are required to explain the observationally-inferred spatial flatness of the universe. Due to their extreme precision or intuitive "unlikeliness," these initial conditions are thought to be unduly special, such that many cosmologists have felt that the initial conditions themselves are in need of explanation and, moreover, present a significant conceptual problem for the HBB model.

Although physical fine-tuning could be interpreted in a variety of ways, cosmologists typically understand it to mean that observationally-required initial conditions are in some sense unlikely (Smeenk, 2013; McCoy, 2015). In order to substantiate this interpretation, one must show that initial conditions in the HBB model which reproduce present conditions are in fact unlikely. This task presupposes that there is a justifiable way of assessing the likelihoods of cosmological models (Gibbons et al., 1987; Hawking and Page, 1988). Many arguments found in the cosmological literature, however, rely on ad hoc, unjustified likelihood measures. Gibbons et al. (1987) propose a "natural" measure (hence the GHS measure) on the set of FRW spacetimes (with matter contents represented by a scalar field) as a natural and justified way of evaluating likelihoods. The GHS measure is simply the canonical Liouville measure associated with the phase space of FRW spacetimes when GTR is put into a Hamiltonian formulation and in a precise sense "comes for free" with the phase space.

While I would maintain that the GHS measure cannot be successfully used to make arguments about fine-tuning in cosmology quite generally, I argue here only for its inapplicability to the flatness problem. Some

*Eidyn Research Centre, University of Edinburgh, Edinburgh, UK. email: casey.mccoy@ed.ac.uk

authors (Gibbons and Turok, 2008; Carroll and Tam, 2010) have attempted to make probabilistic arguments, in analogy to familiar probabilistic arguments in statistical mechanics, by making the GHS measure into a probability measure. However, as the total measure of the FRW phase space is infinite, there is no canonical choice of probability measure with which to make probabilistic arguments, a point that has been recognized already by some (Hawking and Page, 1988; Schiffrin and Wald, 2012). Accordingly, any justification of a particular probability measure is completely independent of the justification of the GHS measure—in short, these probability measures are not in any substantive sense the GHS measure. On the other hand, one might try to use the GHS measure by itself to make typicality arguments in analogy to typicality arguments in statistical mechanics (Goldstein, 2012). Carroll in particular advocates this approach and, interestingly, claims that the GHS measure alone tells us that almost all spacetimes are spatially flat (Carroll and Tam, 2010; Remmen and Carroll, 2013; Carroll, forthcoming)—that there is in fact no flatness problem (Hawking and Page (1988, 803-4) and Coule (1995, 468) suggest the same). Carroll's claim, however, rests on a subtle mistake in interpreting typicality. I claim, on the contrary, that the GHS measure cannot tell us anything about likelihood without substantive additional assumptions such as those made in statistical mechanics, e.g. a partition of phase space into "macroproperties" or similar. These necessary assumptions, however, are doubtfully justifiable in the cosmological context. Thus I ultimately conclude that the GHS measure cannot be used to clarify the nature of fine-tuning in cosmology.

1 The Gibbons-Hawking-Stewart Measure

An adequate view of what the GHS measure is and can do relies on understanding the details of how it is introduced. For this reason I develop here the measure with considerably more care than other accounts in the literature, which tend to jump straight to a Lagrangian or Hamiltonian formulation of GTR without elucidating the geometrical origin of their variable choices and the relations between physical parameters.

My starting point is the initial value formulation of GTR, in which the "position" initial data of spacetime are represented by the spatial metric h_{ab} on a spacelike Cauchy surface Σ and the "momentum" initial data by the extrinsic curvature π_{ab} (Wald, 1984; Malament, 2012). FRW spacetimes are spacetimes with homogeneous and isotropic spacelike hypersurfaces, so one can foliate the spacetimes by a one-parameter family of these spacelike hypersurfaces Σ_t that are orthogonal to a smooth, future-directed, twist-free, unit timelike field ξ^a on M, where I define $\xi^a = \nabla^a t$.For FRW spacetimes the extrinsic curvature of an initial data surface Σ_t is Hh_{ab}, where H is the so-called Hubble parameter.Thus the initial data for an FRW spacetime are completely represented by two objects: (1) the spatial metric h_{ab} and (2) the Hubble parameter H associated with a spatial hypersurface Σ.

The space of initial data is therefore the product of the set of homogeneous and isotropic Riemannian manifolds Σ (with metric h_{ab}) and the set of (real-valued) Hubble parameters H. Homogeneous and isotropic Riemannian manifolds have constant curvature κ. Complete, connected Riemannian manifolds of constant sectional curvature are called space forms. It is a theorem that every simply-connected three-dimensional space form is isometric to the sphere $S^3(\sqrt{(1/\kappa)})$ if $\kappa > 0$, $\mathbf{R}^3$ if $\kappa = 0$, or the hyperbolic space $H^3(\sqrt{(1/\kappa)})$ if $\kappa < 0$ (Wolf, 2010). The standard metrics on each of these manifolds is understood to be the metric induced on them by embedding them in $\mathbf{R}^4$. Every Σ is therefore isometric to one of these three classes of space forms. Spaceforms of each of the three kinds are moreover homothetic, i.e. they are isometric up to the square of a scale factor a (McCabe, 2004). Accordingly one has the means to represent curvature κ as a function of the scale factor; in particular, for any Σ, $a^2\kappa$ is some constant k. Hence one can set any spatial metric $h_{ab} = a^2\gamma_{ab}$, where γ_{ab} is the standard metric on the appropriate space form. This is useful in the initial value formulation of FRW spacetimes because all time dependence of h_{ab} is thereby located solely in

the scale factor rather than in the radius of curvature of the space form.

The Einstein equation reduces to two constraint equations and two evolution equations in the initial value formulation (Geroch, 1972):

$$\mathcal{R} - (\pi_a{}^a)^2 + \pi_{ab}\pi^{ab} = -16\pi T_{ab}\xi^a\xi^b; \tag{1}$$

$$D_c\pi_a{}^c - D_a\pi_c{}^c = 8\pi T_{mr}h^m{}_a\xi^r; \tag{2}$$

$$\pounds_\xi(\pi_{ab}) = 2\pi_a{}^c\pi_{cb} - \pi^c{}_c\pi_{ab} + \mathcal{R}_{ab} - 8\pi h_a{}^m h^n{}_b(T_{mn} - \frac{1}{2}Th_{mn}); \tag{3}$$

$$\pounds_\xi(h_{ab}) = 2\pi_{ab}, \tag{4}$$

where $\mathcal{R}$ is the Ricci scalar of Σ, $\mathcal{R}_{ab}$ is the Ricci tensor of Σ, and D_a is the derivative operator on Σ. For FRW spacetimes, these equations simplify to the following three (the second equation from above is trivial since π_{ab} does not vary across Σ:

$$\mathcal{R} - 6H^2 = -16\pi\rho; \tag{5}$$

$$\dot{H}h_{ab} = \Big(- H^2 - \frac{4\pi}{3}(\rho + 3p)\Big)h_{ab}; \tag{6}$$

$$\dot{h}_{ab} = 2Hh_{ab}, \tag{7}$$

where ρ is the energy density and p the pressure of the matter. The first two equations are known as the Friedman equations. Since $h_{ab} = a^2\gamma_{ab}$, $\dot{h}_{ab} = 2a\dot{a}\gamma_{ab}$, and $2Hh_{ab} = 2Ha^2\gamma_{ab}$, it follows from the third equation above that

$$H = \frac{\dot{a}}{a}, \tag{8}$$

which is the usual definition of the Hubble parameter H. To simplify matters somewhat and to make contact with the literature, I shall henceforth take the matter contents of spacetime to be a scalar field ϕ in a potential V which evolves according to the coupled Einstein-Klein Gordon equation.[1] Then one has the following equations of motion (Hawking and Page, 1988, 790):

$$\mathcal{R} - 6H^2 = -16\pi\Big(\frac{1}{2}\dot{\phi}^2 + V(\phi)\Big) \tag{9}$$

$$\dot{H} = -H^2 - \frac{8\pi}{3}\Big(\frac{1}{2}\dot{\phi}^2 - V(\phi)\Big) \tag{10}$$

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0, \tag{11}$$

where V' is the derivative of the potential with respect to ϕ.[2] (The third equation can be derived from the previous two, and so is in fact redundant.)

For FRW spacetimes the spatial Ricci scalar is $\mathcal{R} = -6\kappa$. As noted before, one can cast κ in terms of the scale factor and a constant k: $\kappa = k/a^2$. By using the scale factor a to replace κ, one has introduced a constant k which has no physical significance beyond identifying whether the space form is flat, positively-curved, or negatively-curved. One therefore usually takes equivalence classes of curves according to these three cases and chooses $k = +1$, 0, and -1 as representatives. Then one may write $\mathcal{R} = -6k/a^2$, so that one finally has Friedman's equation in its usual form (for a scalar field in a potential):

$$\Big(\frac{\dot{a}}{a}\Big)^2 = \frac{8\pi}{3}\Big(\frac{1}{2}\dot{\phi}^2 + V(\phi)\Big) - \frac{k}{a^2}. \tag{12}$$

[1] The scalar field is meant to be the inflaton, the field that drives inflation in the early universe.

[2] If our interest were solely in assessing the HBB model's fine-tuning, one could do the following analysis for perfect fluid matter contents. The results would be qualitatively similar however, as shown by Carroll and Tam (2010, §4.2).

The foregoing indicates that our FRW initial data h_{ab} and π_{ab} are equivalently representable in the space $\{a, \dot{a}, \phi, \dot{\phi}, k\}$. This space is not the space of initial data, however, since the previous equation is a constraint that must be satisfied by initial data. One must also keep in mind that k is an index for three separate copies of the space $\{a, \dot{a}, \phi, \dot{\phi}\}$. There is no continuous path between the three spaces.

Have identified the relevant spaces for representing FRW space forms, I next put the theory into a Hamiltonian formulation (Wald, 1984, Appendix E) in order to obtain a symplectic structure and, hence, the canonical measure. I begin with the Lagrangian for our theory of FRW spacetimes with a scalar field as the matter contents, where I have re-introduced the lapse function N as a Lagrange multiplier:

$$\mathcal{L} = \sqrt{-g}\Big(\frac{R}{16\pi} + \frac{1}{2N^2}\dot{\phi}^2 - V(\phi)\Big). \tag{13}$$

In terms of the variables I have chosen, this is

$$\mathcal{L} = -\frac{1}{8\pi}(\frac{3}{N}a\dot{a}^2 - 3Na^3\frac{k}{a^2}) + \frac{1}{2N}a^3\dot{\phi}^2 - Na^3V(\phi), \tag{14}$$

in agreement with (Hawking and Page, 1988; Gibbons and Turok, 2008; Carroll and Tam, 2010). The momenta of a and ϕ are

$$p_a \equiv \frac{\partial\mathcal{L}}{\partial\dot{a}} = \frac{-3a\dot{a}}{4\pi N}; \qquad\qquad p_\phi \equiv \frac{\partial\mathcal{L}}{\partial\dot{\phi}} = \frac{a^3\dot{\phi}}{N}. \tag{15}$$

The Hamiltonian on this phase space is

$$\mathcal{H} = p_a\dot{a} + p_\phi\dot{\phi} - \mathcal{L} = N\Big(-\frac{2\pi p_a^2}{3a} + \frac{p_\phi^2}{2a^3} + a^3V(\phi) - a^3\frac{3}{8\pi}\frac{k}{a^2}\Big), \tag{16}$$

from which one recovers (after setting $N = 1$) our constraint (the Friedman equation) as the Hamiltonian constraint C:

$$C \equiv -\frac{2\pi p_a^2}{3a} + \frac{p_\phi^2}{2a^3} + a^3V(\phi) - a^3\frac{3}{8\pi}\frac{k}{a^2} = 0. \tag{17}$$

The phase space γ of our system is thus the four-dimensional space $\{a, p_a, \phi, p_\phi\}$ equipped with the canonical symplectic form

$$\omega_{p_a,a,p_\phi,\phi} = \mathrm{d}p_a \wedge \mathrm{d}a + \mathrm{d}p_\phi \wedge \mathrm{d}\phi. \tag{18}$$

The dynamically accessible phase space points are constrained to be on the three-dimensional hypersurface C. Thus it would be inappropriate to use ω for constructing a canonical volume measure on phase space. One can, however, pull the symplectic form back onto the constraint surface by first solving the constraint for p_ϕ:[3]

$$p_\phi = a^3\Big(\frac{4\pi}{3}\frac{p_a^2}{a^4} + \frac{3}{4\pi}\frac{k}{a^2} - 2V(\phi)\Big)^{1/2}. \tag{19}$$

Following Carroll and Tam, I also switch coordinates from p_a to H, so that

$$p_\phi = a^3\Big(\frac{3}{4\pi}(H^2 + k/a^2) - 2V(\phi)\Big)^{1/2} \tag{20}$$

[3]The scalar field can have positive or negative momentum, so strictly speaking there should be a ± in the following equation. The reader is welcome to annotate the equations that follow.

and

$$\mathrm{d}p_a = -\frac{3}{4\pi}\Big(2aH\mathrm{d}a + a^2\mathrm{d}H\Big). \tag{21}$$

The differential of p_ϕ is then

$$\mathrm{d}p_\phi = \frac{(3/4\pi)a^3H\mathrm{d}H - a^3V'\mathrm{d}\phi + 6a^2((3H^2 + 2k/a^2)/8\pi - V)\mathrm{d}a}{((3/4\pi)(H^2 + k/a^2) - 2V)^{1/2}}. \tag{22}$$

Substituting these into ω then gives the pullback of the symplectic form onto C. The result is the following (pre-symplectic) differential form:

$$\omega_{a,H,\phi} = \Theta_{Ha}(\mathrm{d}H \wedge \mathrm{d}a) + \Theta_{H\phi}(\mathrm{d}H \wedge \mathrm{d}\phi) + \Theta_{a\phi}(\mathrm{d}a \wedge \mathrm{d}\phi), \tag{23}$$

where

$$\Theta_{Ha} = -\frac{3}{4\pi}a^2; \tag{24}$$

$$\Theta_{H\phi} = \frac{(3/4\pi)a^3H}{((3/4\pi)(H^2 + k/a^2) - 2V)^{1/2}}; \tag{25}$$

$$\Theta_{a\phi} = \frac{6a^2((3H^2 + 2k/a^2)/8\pi - V)}{((3/4\pi)(H^2 + k/a^2) - 2V)^{1/2}}. \tag{26}$$

This form is not symplectic (it is degenerate), so one cannot construct a natural volume measure on C. Ideally, the "real" phase space of our system would be given by "solving the dynamics," and then taking equivalence classes of phase points that are part of the same trajectory. In this way one would obtain the space of motions, onto which one could then pull back the degenerate form to obtain a new symplectic form (of degree two less than ω) and construct a canonical measure. This is quite complicated in general due to the differential equation that must be solved. The usual approach to take instead is to set H to some value H_* in the differential form and define their measure accordingly, i.e. set

$$\mathrm{d}\Omega = \omega_{a,H,\phi}|_{H=H_*} = \Theta_{a\phi}|_{H=H_*}\mathrm{d}a\mathrm{d}\phi. \tag{27}$$

One may do this because surfaces of constant Hubble parameter in phase space are transverse to temporal evolution, and the measure is preserved under translation of these surfaces along the Hamiltonian flow. Finally, one may naturally define the GHS measure μ_{GHS} on Lebesgue measurable sets U by

$$U \mapsto \int_U \mathrm{d}\Omega = -6\int_U a^2\frac{(3H_*^2 + 2k/a^2)/8\pi - V}{((3/4\pi)(H_*^2 + k/a^2) - 2V)^{1/2}}\mathrm{d}a\mathrm{d}\phi. \tag{28}$$

This expression of the GHS measure is equivalent to those derived in (Carroll and Tam, 2010; Schiffrin and Wald, 2012).[4]

[4]There are some complications with the $k = 1$ case. See (Schiffrin and Wald, 2012, 8) for the details. I have however chosen not to set $8\pi G = 1$, but rather maintained consistency with the rest of this dissertation's use of "geometrical units" by only setting $G = 1$. Gibbons et al. (1987) use a simplifying, but less transparent coordinate choice. They also choose to investigate only the special case where $V = m^2\phi^2/2$. It can be shown with some work that their expression is equivalent to this one as well with this potential.

2 The Flatness Problem

The GHS measure clearly diverges for large scale factors, a point originally recognized by Gibbons et al. (1987, 745); it also converges to 0 for small scale factors. Due to the divergence, one may readily say that, given any choice of Hubble parameter H_*, almost all spacetimes will have a "large" scale factor. More precisely, pick any scale factor a_*; the set of spacetimes with $a < a_*$ is a negligible set: the total measure of this set is finite whereas the total measure of its complement is infinite. What is the significance of this fact about the GHS measure, specifically for the flatness problem?

Hawking and Page (1988, 803-4) suggest the following:

> "Thus for arbitrarily large expansions (and long times), and for arbitrarily low values of the energy density, the canonical measure implies that almost all solutions of the Friedmann-Robertson-Walker scalar equations have negligible spatial curvature and hence behave as $k = 0$ models. In this way a uniform probability distribution in the canonical measure would explain the flatness problem of cosmology..."

By "arbitrarily large expansions" (and "arbitrarily low values of energy density"), they appear to mean the following. Pick any arbitrary a_* (and any arbitrary ϕ_*).[5] According to the GHS measure almost all spacetimes have $a > a_*$ (and $\phi > \phi_*$), or, equivalently, the spacetimes with $a < a_*$ (and $\phi < \phi_*$) compose a negligible set. Furthermore, since this holds for any choice of a_*, one may infer that almost all spacetimes are arbitrarily close to having $\kappa = 0$ (since $\kappa = k/a^2$) in exactly the same sense. It is perhaps somewhat misleading to say that curved FRW spacetimes with large scale factors "behave as $k = 0$ models;" the curvature does not change in such models. It is, however, surely false to say that a "uniform probability distribution" with respect to the GHS measure would explain the flatness problem of cosmology. There is in fact no such uniform probability distribution, since the GHS measure is not finite. Moroever, there is also no canonical probability distribution ρ at all which would make $U \mapsto \int_U \rho \mathrm{d}\Omega_{GHS}$ into a probability measure—one has to make a choice in order to obtain a probability measure in the case of infinite total measure, a choice which appears completely arbitrary in this context.

Carroll and Tam (2010, 14) invite us to consider the question in more "physically transparent" terms by looking at the curvature κ, which I previously exchanged in favor of the scale factor a when deriving the GHS measure. One can recast the scale factor a as the curvature κ using the relation from before, namely $\kappa = k/a^2$. (Note especially that this switch maps the entire set of scale factors for the $k = 0$ case to the single point $\kappa = 0$.) One then defines the GHS measure (at least for curved FRW spacetimes) by the map

$$U \mapsto \int_U \mathrm{d}\Omega = -6 \int_U \frac{1}{|\kappa|^{5/2}} \frac{(3H_*^2 + 2\kappa)/8\pi - V}{((3/4\pi)(H_*^2 + \kappa) - 2V)^{1/2}} \mathrm{d}\kappa \mathrm{d}\phi. \tag{29}$$

It is clear that the measure diverges for small values of curvature, i.e. curvatures close to flat, due to the curvature term in the denominator. This is pointed out by Carroll and Tam (2010, 15). They suggest the following interpretation of this fact:

> "Considering first the measure on purely Robertson-Walker cosmologies (without perturbations) as a function of spatial curvature, there is a divergence at zero curvature. In other words, curved [FRW] cosmologies are a set of measure zero—the flatness problem, as conventionally understood, does not exist."

[5] Gibbons and Turok (2008, 6) point out that ϕ is always bounded given H_*, so it is not really necessary to pick an arbitrary ϕ_*.

As stated these claims are highly suspect.

Firstly, Carroll and Tam assert that all values of their curvature coordinate Ω_k (essentially equivalent to κ) can be integrated over. While this is perhaps true, portraying the phase space in terms of curvature is misleading. For curved FRW spacetimes, it is true that the measure diverges for small values of curvature κ, as I indicate above and as Hawking and Page suggest in the passage from their paper quoted above. The recast measure, however, is infinite *at* zero curvature because the entire set of $k = 0$ scale factors is mapped to $\kappa = 0$. The GHS measure diverges for large scale factors in the case of flat FRW spacetimes just as it does for curved FRW spacetime. Thus it is misleading to describe a "divergence at zero curvature;" there is nothing special going on in flat FRW spacetimes (at least in this respect).[6]

Secondly (and relatedly), curved FRW spacetimes are clearly not a set of measure zero—at least according to the GHS measure. The initial data of FRW spacetimes is representable in the space $\{a, \dot{a}, \phi, \dot{\phi}, k\}$. The curvature constant k serves as an index for *three different phase spaces*, each of which has an infinite total measure—even after taking into account constraints and choosing a hypersurface in the constraint surface according to GHS's procedure. The unboundedness of the total phase space measure for each kind of FRW spacetime is due, again, to the unbounded range of the scale factor Schiffrin and Wald (2012, 11).[7] This is quite plain when one expresses the GHS measure in terms of the scale factor. Transforming to the curvature coordinate κ should not change the fact that the total measure of each phase space is infinite. So, while it is true that the GHS measure attributes infinite measure to flat FRW spacetimes (as Carroll and Tam appear to recognize), it also does so both to positively curved FRW spacetimes and to negatively curved spacetimes. Therefore it is false that the curved FRW cosmologies are a set of measure zero according to the GHS measure; hence one cannot conclude on this basis that the flatness problem does not exist.

One might try to rescue Carroll and Tam's claim about the flatness problem by interpreting flatness more broadly, namely by including "nearly flat" curved spacetimes. This requires specifying what the set of "nearly flat" curved spacetimes is to be, e.g. a specification of the set of spacetimes with curvature less than some κ_* (at some time corresponding to Hubble parameter H_*). Almost all spacetimes will have a "small" curvature κ in comparison to this curvature κ_*. In other words, the set of spacetimes with $\kappa > \kappa_*$ is a negligible set. Since our universe's spatial curvature is thought to be "nearly flat," i.e. it should be less than κ_* (whatever it is), it follows from this argument that our universe is actually typical, contra what is assumed in the flatness problem. Unfortunately this argument does not follow from the GHS measure alone, since one had to make an independent choice in choosing κ_*, a choice that is not natural in any clear sense whatever. Furthermore, it is doubtful that there is any reasonable argument to justify a choice of κ_*—an explication of "close to flat" in the context of FRW models; it appears to be a completely arbitrary choice.

Here is a slightly different tack into the same stiff headwind. Suppose κ_* is the (non-zero) spatial curvature of our universe at the present time. The GHS measure can be used to infer that almost all spacetimes with the same Hubble parameter will have flatter spatial curvatures. In such circumstances, one might be inclined to wonder "Why is my universe's spatial curvature so large? It seems like it ought to be much smaller if my universe is typical!" On this line of thought, it seems like one actually has a curvature problem rather than a flatness problem. Of course one would say this for any κ_* whatsoever, regardless of its magnitude,

[6] Carroll and Tam appear to equivocate several times between there being a divergence *at* $\kappa = 0$ and the measure diverging *as* $\kappa \to 0$: "The integral diverges near [$\kappa = 0$], which is certainly a physically allowed region of parameter space" (Carroll and Tam, 2010, 17); "The measure diverges on flat universes" (Carroll and Tam, 2010, 28).

[7] Besides in (Schiffrin and Wald, 2012), this fact is correctly pointed out in (Gibbons et al., 1987; Hawking and Page, 1988). While Carroll and Tam (2010, 20-1) observe that "this divergence was noted in the original GHS paper, where it was attributed to 'universes with very large scale factors' due to a different choice of variables," they object to this as an interpretation: "This is not the most physically transparent characterization, as any open universe will eventually have a large scale factor." For this reason they exchange the scale factor for curvature; it is not clear, however, how this characterization is more physically transparent since it amounts to the same thing.

so it is not clear how one would ever be in the position to be satisified with one's curvature in an FRW universe—at least insofar as one expects things in our universe to be typical (in accord with Copernican principle-style reasoning). No matter. The measure suggests this question. What is the answer?

The answer is that the curvature depends on the actual dynamical history of the universe, and so it has no explanation within the context of the HBB model (apart from one depending on an initial condition). That answer may be unsatisfying, but the question is a bad one anyway, driven by misleading intuitions. There is no such thing as a typical FRW spacetime, and the GHS measure is not going to explain why the universe's curvature is what it is. This kind of thinking is clearly motivated by supposing that the GHS measure can be used as a likelihood measure, as Carroll and Tam clearly do:

> "When we consider questions of fine-tuning, however, we are comparing the real world to what we think a randomly-chosen history of the universe would be like" (Carroll and Tam, 2010, 11).

Some popular, specious conceptions (in physics and beyond) of statistical mechanics encourage this line of thought. Putatively successful typicality arguments in statistical mechanics (Goldstein, 2012) depend, however, not only on having a phase space measure, but also on both the dynamics of the system and on a specification of macroproperties or macrostates (defined as regions of phase space) (Frigg, 2009; Frigg and Werndl, 2012). Accordingly, any claim of fine-tuning in FRW spacetimes on the sole basis of the GHS measure (which does at least incorporate the FRW dynamics) is bound to miss the mark without additional assumptions (such as a well-motivated standard of flatness).

Gibbons and Turok (2008) take a different approach from Carroll and Tam. They correctly observe that universes with large scale factors are universes with small spatial curvatures. They then claim that the scale factor is neither "geometrically meaningful" nor "physically observable" and therefore propose to identify all the "indistinguishable" nearly flat spacetimes on the surface identified by H_*.[8] They do so by effectively choosing a "cutoff" curvature κ_* and throwing out all the spacetimes with curvatures smaller than it. The advantage to doing this is that the total measure of FRW spacetimes with curvatures larger than κ_* is finite, so that one can then define a probability measure in a natural way.

The disadvantage is that this makes no sense. Carroll and Tam (2010, 20) comment, "to us, this seems to be throwing away almost all the solutions, and keeping a set of measure zero. It is true that universes with almost identical values of the curvature parameter will be physically indistinguishable, but that doesn't affect the fact that almost all universes have this property." Indeed, doing what Gibbons and Turok do is throwing away almost all the solutions (although the remaining set has finite measure, not measure zero as Carroll and Tam claim). They are also right to point out that if nearly flat universes are physically indistinguishable, so are "nearly-κ" universes for almost any κ. Gibbons and Turok do not throw out these universes however (else they would not have been left with any universes at all). Their justification for an additional assumption therefore fails.

Ironically, Carroll and Tam make essentially the same error as Gibbons and Turok, by identifying the flat and nearly flat spacetimes. Instead of throwing out all the flat and nearly flat spacetimes like the latter pair, however, the former pair throws out the complement of the flat and nearly flat spacetimes by assigning them zero measure. They then triumphantly conclude that all FRW spacetimes are essentially flat! Carroll and Tam propose to tame the remaining divergence in the GHS measure by regularizing the integral, in effect making the measure finite. The problem with doing this is that, since the GHS measure is not finite,

[8] It is not clear what they mean by "geometrically meaningful." The scale factor is clearly geometric in the relevant sense, since it relates spaceforms of the same kind by scalings. It is moreover physically meaningful because space is expanding (or contracting) in FRW spacetimes. The precise value of a does not matter, as it can be re-scaled, but that does not undermine its meaningfulness. It is also unclear how the fact that a is physically unobservable should matter, since most features of spacetime are not observable, e.g. the metric g, the spatial curvature κ, etc. The physically relevant content of these, including the scale factor, can be inferred from observations and appropriate assumptions.

regularizing the measure makes it no longer the GHS measure, in which case any justification the measure had by its "naturalness" is lost since a choice was made.[9] In short, one may as well have just assumed the probability distribution they end up with from the very beginning.Their stated justification for this move is pragmatic: "This non-normalizability is problematic if we would like to interpret the measure as determining the relative fraction of universes with different physical properties" (Carroll and Tam, 2010, 17). However this is obviously an inadequate justification for the propriety of their measure.

References

Albrecht, Andreas, and Paul Steinhardt. "Cosmology for Grand Unified Theories with Radiatively Induced Symmetry Breaking." *Physical Review Letters* 48: (1982) 1220–1223.

Belinsky, Vladimir, Leonid Grishchuk, Isaak Khalatnikov, and Yakov Zeldovich. "Inflationary Stages in Cosmological Models with a Scalar Field." *Physics Letters B* 155: (1985) 232–236.

Carroll, Sean. "In What Sense Is the Early Universe Fine-Tuned?" In *Time's Arrows and the Probability Structure of the World*, edited by Barry Loewer, Brad Weslake, and Eric Winsberg, Cambridge, MA: Harvard University Press, forthcoming.

Carroll, Sean, and Heywood Tam. "Unitary Evolution and Cosmological Fine-Tuning." ArXiv Eprint, 2010. `http://arxiv.org/abs/1007.1417`.

Coule, David. "Canonical measure and the flatness of a FRW universe." *Classical and Quantum Gravity* 12: (1995) 455–469.

Dicke, Robert, and Jim Peebles. "The Big Bang Cosmology—Enigmas and Nostrums." In *General Relativity: An Einstein Centenary Survey*, edited by Stephen Hawking, and Werner Israel, Cambridge: Cambridge University Press, 1979, chapter 9, 504–517.

Frigg, Roman. "Typicality and the Approach to Equilibrium in Boltzmannian Statistical Mechanics." *Philosophy of Science* 76: (2009) 997–1008.

Frigg, Roman, and Charlotte Werndl. "Demystifying Typicality." *Philosophy of Science* 79: (2012) 917–929.

Geroch, Robert. "General Relativity.", 1972. Unpublished lecture notes.

Gibbons, Gary, Stephen Hawking, and John Stewart. "A natural measure on the Set of all Universes." *Nuclear Physics B* 281: (1987) 736–751.

Gibbons, Gary, and Neil Turok. "Measure problem in cosmology." *Physical Review D* 77: (2008) 1–12.

Goldstein, Sheldon. "Typicality and Notions of Probability in Physics." In *Probability in Physics*, edited by Yemima Ben-Menahem, and Meir Hemmo, Berlin: Springer Verlag, 2012, chapter 4, 59–71.

Guth, Alan. "Inflationary universe: A possible solution to the horizon and flatness problems." *Physical Review D* 23, 2: (1981) 347–356.

[9]Carroll more recently has conceded the artificiality of regularizing: "Earlier attempts to regularize the measure, for example by considering an ϵ-neighborhood around the zero-curvature Hamiltonian constraint surface (Carroll and Tam, 2010) or by identifying universes with similar curvatures (Gibbons and Turok, 2008) have not proven satisfactory" (Remmen and Carroll, 2013, 7). He remains convinced, however, that almost all FRW spacetimes are "nearly flat:" "we should throw all of the others away and deal with flat universes," (Carroll, forthcoming, 19), developing a measure on just these spacetimes in a later paper (Remmen and Carroll, 2014).

Hawking, Stephen, and Don Page. "How Probable is Inflation?" *Nuclear Physics B* 298: (1988) 789–809.

Linde, Andrei. "A New Inflationary Universe Scenario: A Possible Solution of the Horizon, Flatness, Homogeneity, Isotropy, and Primordial Monopole Problems." *Physics Letters B* 108: (1982) 389–393.

———. "The inflationary universe." *Reports on Progress in Phyics* 47: (1984) 925–986.

Malament, David. *Topics in the Foundations of General Relativity and Newtonian Gravity Theory*. Chicago: University of Chicago Press, 2012.

McCabe, Gordon. "The structure and interpretation of cosmology: Part I—general relativistic cosmology." *Studies in History and Philosophy of Modern Physics* 35: (2004) 549–595.

McCoy, Casey. "Does inflation solve the hot big bang model's fine-tuning problems?" *Studies in History and Philosophy of Modern Physics* 51: (2015) 23–36.

Remmen, Grant, and Sean Carroll. "Attractor solutions in scalar-field cosmology." *Physical Review D* 88: (2013) 1–14.

———. "How many *e*-folds should we expect from high-scale inflation?" *Physical Review D* 90: (2014) 1–14.

Schiffrin, Joshua, and Robert Wald. "Measure and probability in cosmology." *Physical Review D* 86: (2012) 1–20.

Smeenk, Chris. "Philosophy of Cosmology." In *The Oxford Handbook of Philosophy of Physics*, edited by Robert Batterman, Oxford: Oxford University Press, 2013, chapter 17, 607–652.

Wald, Robert. *General Relativity*. Chicago: University of Chicago Press, 1984.

Wolf, Joseph. *Spaces of Constant Curvature*. Providence, RI: AMS Chelsea Publishing, 2010, 6th edition.

Invariance, Interpretation, and Motivation

Thomas Møller-Nielsen

July 2016

[Forthcoming in *Philosophy of Science (2016 Proceedings)*.]

Abstract

In this paper I assess the 'Invariance Principle', which states that only quantities that are invariant under the symmetries of our theories are physically real. I argue, contrary to current orthodoxy, that the variance of a quantity under a theory's symmetries is not a sufficient basis for interpreting that theory as being uncommitted to the reality of that quantity. Rather, I argue, the variance of a quantity under symmetries only ever serves as a motivation to refrain from any commitment to the quantity in question. In the process of this discussion, I address the related but importantly distinct issue of when symmetries can be said to prompt a mathematical reformulation of the relevant theory.

1 Introduction

Take the *Invariance Principle* to be the principle that only quantities that are invariant under the symmetries of our theories are physically real.[1] It is a doctrine with a distinguished pedigree: acclaimed theorists as diverse as the physicist Paul Dirac, the mathematician Hermann Weyl, and the philosopher Robert Nozick were all apparent signatories during their respective lifetimes.[2] *Prima facie*, however, it is something of a mystery as to how and why the principle is supposed to work. Nevertheless, there appear to be at least some uncontroversial cases where it—or something very close to it—does work.

One such example can be found in Newtonian Gravitation Theory (NGT), i.e., the theory comprising Newton's three laws, plus his inverse square gravitational law, governing the behaviour of point particles in Newtonian spacetime. As is well known, this theory is *Galilean invariant*. This implies, among other things, that if one takes any solution to NGT and "boosts" it—that is, uniformly alters the absolute velocity of each point particle by the same amount throughout its history—one will invariably get back a solution to NGT. Boosts, in other words, are a *symmetry* of NGT: they are transformations that invariably map solutions of the theory to solutions.

[1] I draw the term from Saunders (2007). Compare also Dasgupta's (forthcoming) "symmetry-to-reality inference".

[2] See, e.g., Dirac (1930, vii), Weyl (1952, 132), and Nozick (2001, 82).

Which quantity varies under this particular symmetry? The answer is obvious: absolute velocity. Thus, according to the Invariance Principle, we should conclude that absolute velocity is not a genuine physical quantity. Conversely, which quantities are invariant under this particular symmetry? Again, the answer is obvious: relative (inter-particle) distance and velocity, temporal intervals, and absolute acceleration. Thus, according to the Invariance Principle, we should conclude that NGT's boost symmetry does not threaten these quantities' status as genuinely physical.

As it turns out, one can successfully purge Newtonian theory of the spacetime structure required to make absolute velocity a physically meaningful quantity. More specifically, one can move to *Galilean spacetime*. (Sometimes also called "Neo-Newtonian spacetime".)[3] Here, the Newtonian posit of persisting points of absolute space—persisting points which, crucially, allow for the notion of absolute velocity to be physically meaningful—is done away with, but an *affine structure* is nevertheless preserved, which defines the "straight" or force-free (inertial) paths through spacetime. Absolute velocity is therefore not a physically meaningful quantity in Galilean spacetime, as it is in Newtonian spacetime. Nevertheless, all other Newtonian notions, including the notion of absolute acceleration, remain well-defined in Galilean spacetime. To the extent that one opts for Galilean over Newtonian spacetime, then, one has excised an ostensibly odious piece of theoretical structure from NGT.

Three important caveats are worth noting, however. First, and most obviously, none of this is to say that Newtonian theory set in Galilean spacetime is therefore the true and complete theory of the world. (It isn't.) Second, nor is this to say that by moving to Galilean spacetime one has thereby purged Newtonian theory of all its "variant" structure. (One hasn't. The symmetry group of Newtonian theory is actually wider than the Galilean group: it has additional symmetries.)[4] Third, nor is this even to say that the invariant quantities one ends up with following such an application of the Invariance Principle will invariably be preserved in future theories. (For instance, there is no notion of "relative spatial distance" *simpliciter* in special relativity.) Given all of these caveats, however, one might well ask: What good is the Invariance Principle, exactly? What purpose, in particular, does it serve?

As I see it—and, I take it, as many other contemporary theorists also see it—the purpose of the Invariance Principle is essentially *comparative*. That is, it is simply supposed to lead you to a *better theory*—or a better interpretation, or characterisation, of the same theory—than the one you started with. To take the case at hand: Newtonian theory set in Galilean spacetime is a better theory than Newtonian theory set in Newtonian spacetime. It is a theory which possesses all of the theoretical virtues of its rival, but lacks any apparent ontological commitment to the unwanted variant quantity in question.

In summary, the Galilean invariance of NGT, in conjunction with the Invariance Principle, is supposed to indicate that neither absolute velocity nor

[3] See, e.g., Earman (1989, §2.4).

[4] See, e.g., Knox (2014). I discuss this point further in Section 4 below.

any corresponding persisting points of absolute space are genuinely real. Now to lay my cards on the table: I actually think that something *very close* to this general kind of inference—that is, from the variance of a quantity under symmetries to that quantity's nonreality—is legitimate. The devil, however, is in the details. In particular, I don't believe that the *mere* Galilean invariance of NGT is enough to establish absolute velocity's nonreality. And in general, I don't believe that the *mere* variance of a quantity under symmetries is enough to establish that quantity's nonreality. These beliefs, as far as I can determine, put me in the minority camp in the contemporary philosophical literature on symmetries. Nevertheless, I think they are correct beliefs—and they are precisely the ones that I will attempt to argue for in the remainder of this paper.

2 Interpretational vs Motivational

In arguing for the above claims, it will prove extremely useful first to distinguish between two very different ways of thinking about symmetries.

Close cousins of the distinction that I have in mind have already been drawn in the literature. Thus, Greaves and Wallace write:

> There is a widespread consensus that two states of affairs related by a symmetry transformation are really just the same state of affairs differently described. That is, if two mathematical models of a physical theory are related by a symmetry transformation, then those models represent one and the same physical state of affairs. (Greaves and Wallace 2014, 60)

They continue:

> Although we agree with this consensus [...] even those who do not agree that symmetry-related states of affairs are identical at least agree that they are *empirically indistinguishable* from one another. (Greaves and Wallace 2014, 60, fn 1)

To illustrate the difference between these two ways of thinking about symmetries, consider again the example of boosts in NGT. According to the "widespread consensus" view alluded to, and endorsed by, Greaves and Wallace, boosted models of NGT are to be taken to represent the same physical state of affairs *even when the theory is putatively set in Newtonian spacetime*. In other words, according to this view, one needn't make the move to Galilean spacetime in order not to be committed to absolute velocities; there is a way of understanding boosted models' physical equivalence, and their associated noncommitment to the notion of absolute velocity, prior to making this move.[5]

Things are very different according to the second conception of symmetries described, and rejected, by Greaves and Wallace. According to this view, boosted models of NGT are to be regarded as physically *in*equivalent: they are not to be construed as representing the same physical state of affairs. Instead,

[5]See, e.g., Healey (2007, 114-7), for an endorsement of this view in the Newtonian context.

such models are taken to represent physically distinct scenarios, which differ in what absolute velocity they ascribe to the world's total material content. Nevertheless, such models still represent *empirically indistinguishable* states of affairs: in a Newtonian universe, no experiment could ever help an observer determine what her absolute velocity actually is. Such boosted models therefore represent physically distinct ways for the world to be, albeit ones that are indiscernible on the basis of measurement.[6]

As previously mentioned, this distinction between different ways of thinking about symmetries is close, but not identical, to the one that I want to draw. The key reason why it is not identical is because Greaves and Wallace say nothing to the effect that the person who subscribes to the second conception of symmetries—that is, who believes that symmetry-related models invariably represent empirically indistinguishable, but not necessarily physically equivalent, states of affairs—should still be *motivated to seek* an alternative theory, or an alternative interpretation or characterisation of the same theory, according to which such models do not merely represent empirically indistinguishable scenarios, but rather represent physically equivalent states of affairs.[7] Moreover, I claim, it is precisely this notion of *motivation* which plays a central role in correctly understanding the philosophical significance of symmetries in the general case.[8]

Here, then, is what I take to be the appropriate distinction between these two different ways of thinking about symmetries:

- **Interpretational**: Symmetries allow us to *interpret* theories as being committed solely to the existence of invariant quantities, even in the absence of a metaphysically perspicuous characterisation of the reality which is alleged to underlie symmetry-related models.
- **Motivational**: Symmetries only *motivate* us to find a metaphysically perspicuous characterisation of the reality which is alleged to underlie symmetry-related models, but they do not allow us to interpret that theory as being solely committed to the existence of invariant quantities in the absence of any such characterisation.

The central claim of this paper may now be neatly summarised: the (orthodox) interpretational view is mistaken; the (unorthodox) motivational view is correct.

Drawing the distinction in the way that I have done, however, invites the rather obvious question: What, precisely, is meant by a "metaphysically perspicuous characterisation" of reality? This is the question addressed in the next section.

[6] See, e.g., Maudlin (1993, 192), for an endorsement of this view in the Newtonian context.

[7] Compare (again) Maudlin's (1993, 192) discussion in the Newtonian context.

[8] Note that I do not intend any of this as a criticism of Greaves and Wallace's paper. Indeed, as Greaves and Wallace (2014, 60, fn 1) are careful to remark, the distinction they draw is orthogonal to the central topic of their paper, namely the issue of which symmetries have "direct empirical significance" (i.e., have analogues to Galileo's ship).

3 More on Metaphysical Perspicuity

In intuitive terms, a metaphysically perspicuous characterisation of reality is one which corresponds to, or "limns", reality's structure in some suitably faithful way. To use another common (Platonic) metaphor, a metaphysically perspicuous characterisation of reality is one which "carves nature at its joints". (In comparative terms: a description of reality is *more* metaphysically perspicuous than another precisely to the extent that it corresponds to, or limns, reality's structure *more* faithfully than its rival does.)

As many readers will be aware, such a notion is frequently alluded to, and made use of, in contemporary analytic metaphysics.[9] But metaphysical perspicuity is also, I think, a notion that is reasonably serviceable in physical (rather than "merely metaphysical") contexts. One particularly illustrative example—albeit a slightly misleading one, for reasons that I will soon explain—drawn from physics may plausibly be found in classical electromagnetism.[10] As is well known, this theory may be formulated in two different ways.[11] According to one such formulation, EM_1, the theory is expressed in terms of the Faraday tensor, F_{ab}, satisfying the (Maxwell) equations $\nabla_{[a}F_{bc]} = 0$ and $\nabla_a F^{ab} = J^a$, where J^a is a vector field representing the charge current density. According to the second formulation, EM_2, however, the theory is expressed in terms of the vector potential, A_a, satisfying the equation $\nabla_a\nabla^a A^b - \nabla^b\nabla_a A^a = J^b$.

These two formulations of electromagnetism are related to one another. In particular, any model $\langle M, \eta_{ab}, A_a\rangle$ of EM_2 corresponds to a unique model $\langle M, \eta_{ab}, F_{ab}\rangle$ of EM_1, via the equation $F_{ab} = \nabla_{[a}A_{b]}$. The converse, however, is not true. That is, a typical model of EM_1 does *not* typically correspond to a unique model of EM_2. More specifically, if $\langle M, \eta_{ab}, A_a\rangle$ is a model of EM_2 corresponding to a model $\langle M, \eta_{ab}, F_{ab}\rangle$ of EM_1, then so will any other model of EM_2 $\langle M, \eta_{ab}, A'_a\rangle$, where A'_a is related to A_a by a "gauge transformation" $A'_a = A_a + \nabla_a\chi$, where χ is some smooth scalar field.

It is EM_1 which, I take it, constitutes the metaphysically perspicuous characterisation of this theory. That is, it is the tensor F_{ab} which faithfully represents the fundamental ontology of the theory, namely the electromagnetic field. Not so EM_2. This second formulation may, of course, be useful for various calculational or heuristic purposes. But the key point is that the vector potential A_a *does not directly represent a genuinely real field*: rather, it is merely a mathematically convenient "shorthand" way of characterising and determining the values of the Faraday tensor, which *is* taken to represent the genuine material ontology of the theory.[12] Moreover, it is precisely by construing the vector potential in this

[9] See, e.g., O'Leary-Hawthorne and Cortens (1995, 154-7).

[10] Here and below, I take this theory to be set in Minkowski spacetime. Thus, the spacetime models of this theory are of the form $\langle M, \eta_{ab}\rangle$, where M is a four-dimensional differentiable manifold, and η_{ab} is the Minkowski metric.

[11] For a recent, intriguing study of the relationship between these two different formulations of electromagnetism, see Weatherall (forthcoming). I draw heavily on his discussion over the next couple of paragraphs.

[12] Modulo, that is, certain concerns that arise as a result of the Aharonov-Bohm effect. See, e.g., Healey (2007).

way which plausibly allows us to explain and understand, in a fully transparent way, gauge-symmetry models' physical equivalence in EM_2—namely, for the reason that they are merely notationally distinct ways of representing the same fundamental physical ontology.

As mentioned above, I think this example of metaphysical perspicuity is apt to be slightly misleading, at least when taken on its own. This is because this example might make it seem as though having a metaphysically perspicuous characterisation of the (putative) reality underlying symmetry-related models crucially relies upon one having to *mathematically reformulate* the relevant theory (or at least upon having such a mathematical reformulation already in hand), and in particular upon having to reformulate the theory so as to remove any relevant representational redundancy. However, I think this is incorrect. That is, I believe that one *can*, in fact, be in possession of a metaphysically perspicuous characterisation of the reality underlying symmetry-related models *even in the absence* of any mathematical (re-)formulation of the theory which removes the relevant representational redundancy.

Let me illustrate this point with two simple examples. First, consider the case of *shift symmetry* in NGT. This symmetry is subtly different from the case of boost symmetry, discussed above. Here, instead of uniformly altering the absolute velocity of each particle throughout its history, one enacts a global, time-independent repositioning of all matter in space. Thus, for instance, in the shifted world all of the world's material content will (*prima facie*) be located three metres to the left of where it is in the original world. The basic idea behind the "Leibniz shift" argument—the famous argument associated with this symmetry—is that the substantivalist's admission of points of space as primitive objects (allegedly) has the undesirable consequence of committing her to regarding shifted worlds as physically distinct, yet nevertheless empirically indistinguishable:[13] in intuitive terms, everything would look, feel, taste, touch and sound the same in the two (putatively distinct) shifted worlds, just as in the case of boosted worlds.

It will prove helpful to express all of this in terms of the models of the theory. Thus, take a generic model of NGT to be of the form $\mathcal{M} = \langle M, t_{ab}, h^{ab}, \sigma^a, \rho, \phi \rangle$, where M is a differentiable 4-dimensional manifold, t_{ab} is the temporal metric, h^{ab} is the spatial metric, σ^a is the timelike vector field whose integral curves represent the persisting points of absolute space, and ρ and ϕ represent the matter density and the gravitational potential field respectively.[14] A shift symmetry can then be characterised as the application of the appropriate diffeomorphism (corresponding to a spatial translation) d so as to yield a new model $\mathcal{M}_{static} = \langle M, t_{ab}, h^{ab}, \sigma^a, d^*\rho, d^*\phi \rangle$. It is then alleged that $\mathcal{M}$ and $\mathcal{M}_{static}$ differ precisely

[13] Though see Maudlin (1993), who notes that there is an interesting (epistemological) sense in which shifted worlds in NGT are not indiscernible after all.

[14] Note that the canonical presentations of Newtonian spacetime (e.g., Earman 1989, §2.5) take the affine connection as ideologically primitive. I find such presentations unsatisfactory for historical rather than for philosophical reasons: in particular, it threatens to make the move to Galilean spacetime seem almost trivial, and the associated timelike vector field trivially superfluous. For more on this point, see Pooley (MS, §4.4-§4.5).

insofar as they each represent the world's matter content as being located at distinct places in absolute space. More specifically, such Leibniz-shifted scenarios are alleged to differ precisely with regard to which particular points of space are underlying various parts of the matter fields.

For a second example, consider *diffeomorphism symmetry* in general relativity (GR). Here, similarly, the existence of this symmetry is alleged to commit the substantivalist to a plurality of physically distinct possibilities that are nevertheless empirically indistinguishable. In terms of the models of the theory: taking a generic model of GR to be of the form $\mathcal{M} = \langle M, g_{ab}, T_{ab} \rangle$ and applying an arbitrary diffeomorphism d to yield a new model $\mathcal{M}_{diff} = \langle M, d^*g_{ab}, d^*T_{ab} \rangle$ (where M is again a differentiable 4-dimensional manifold, g_{ab} is the metric tensor, and T_{ab} is the stress-energy tensor which, roughly speaking, represents the model's matter content), the two scenarios represented are alleged to differ with regard to which particular points of the spacetime manifold are underlying various parts of the metric and matter fields.[15]

It is my contention that neither the shift symmetry of NGT, nor the diffeomorphism symmetry of general relativity, by themselves motivate any mathematical reconstrual of the respective theories. This is because I believe there is a perfectly transparent, anti-haecceitist, "modestly structuralist"—but nevertheless fully substantivalist—way of understanding such models' representational equivalence even in the absence of any such mathematical reformulation. On this view, spacetime points are construed as genuinely real, fundamental entities. However, they are "contextually individuated": they are not to be understood as being anything more—or less—than "nodes" in the relational, geometrical structures in which they are embedded. Shifted models in NGT and diffeomorphically-related models in GR are thus to be understood as representing the same physical state of affairs precisely because the exact same pattern of relational, geometrical structures is represented as obtaining in each case. Moreover, this view denies that there are any primitive, singular ("haecceitistic") facts about spacetime points which would even allow for a distinction between shifted or diffeomorphically-related scenarios to be coherently drawn.[16]

Whence the difference, then, between the case of gauge symmetry in electromagnetism on the one hand, and shift and diffeomorphism symmetry in NGT and GR on the other? I think the answer is straightforward. In the latter cases, the models in question are *isomorphic*: they represent worlds which differ at most with regard to which particular objects are playing which qualitative roles, i.e., they represent at most haecceitistically distinct possible worlds. Hence, adopting modest structuralism (which implies anti-haecceitism) about spacetime transparently collapses the number of possibilities represented by these models to one. In the former such case, however, the relevant models are *not* isomorphic—read "literally", gauge-related models of EM_2 assign *qualitatively distinct* arrangements of the vector field over spacetime—hence adopting a modestly structuralist ontology does not by itself collapse the number of represented

[15] For further details see, e.g., Earman (1989, §9).

[16] For further defence of this view—which is sometimes also called *sophisticated substantivalism* in the literature—see, e.g., Saunders (2003), Ladyman (2007), and Pooley (2013).

possibilities to one. In order to transparently understand such models' physical equivalence, then, a mathematical reformulation of the theory is required.

To summarise the claims made thus far: according to the motivational view of symmetries, one is invariably only motivated to regard symmetry-related models as physically equivalent; moreover, one is justified in regarding such models as physically equivalent only insofar as one is in possession of a metaphysically perspicuous characterisation of the reality which is alleged to underlie them. However, it is possible to be in possession of a metaphysically perspicuous characterisation of the reality underlying symmetry-related models even in the absence of a mathematical formulation of the theory which removes the relevant representational redundancy. Such a metaphysically perspicuous characterisation is possible just in case the symmetry-related models in question are isomorphic, or are naturally understood as representing at most haecceitistically distinct possibilities. In brief: symmetry-related, isomorphic models invariably do *not* motivate a mathematical reformulation of the relevant theory (modest structuralism invariably suffices); but symmetry-related, *non*-isomorphic models invariably *do*.[17]

4 In Defence of the Motivational View

Let us return once more to the case of NGT. As alluded to in Section 1, the symmetry group of this theory is quite large. For not only does it include transformations corresponding to global velocity boosts of solutions' matter content, but it also includes transformations corresponding to time-dependent translational accelerations of such content (so long as the gravitational potential field is also appropriately transformed). Thus, read "literally", the symmetries of this theory include transformations that map solutions to solutions that represent physically distinct, but nevertheless empirically indistinguishable, states of affairs in which a given material system is:

1. Force-free and stationary with respect to absolute space.
2. Force-free and moving at constant absolute velocity.
3. Absolutely accelerating under a gravitational force-field.

According to the interpretational conception of symmetries, we may legitimately take all of these symmetry-related solutions to in fact represent the same physical state of affairs—despite the fact that they are naturally understood as representing radically distinct physical situations. Things are very different, however, according to the motivational conception of symmetries. On this view, we are merely *motivated to regard* all such solutions as representing the same physical state of affairs, the motivation arising from the general Occamist principle that, other things being equal, our preferred scientific theories should not allow for solutions that represent physically distinct but nevertheless empirically indistinguishable possible worlds. According to the motivational

[17] See also Pooley (2013, 576-7) and Weatherall (forthcoming) for recent, related arguments to this effect.

view, then (and to repeat slightly), absent a metaphysically perspicuous characterisation of the reality underlying these symmetry-related models, we have no choice but to regard them as representing physically distinct states of affairs.

For our purposes, the crucial thing to note about all of these models is that *none of them are isomorphic*—naturally understood, they do not represent at most haecceitistically distinct possible worlds. According to the criterion laid down in the previous section, then, in order to be able to transparently understand how it could be that such models may be said to represent physically equivalent scenarios, a mathematical reformulation of the theory is required.

As it turns out, such a mathematical reformulation of the theory is possible. In brief, in this reformulation one replaces the vector field σ^a with a new kind of *dynamical* inertial connection ∇^{NC}, with models of the form $\mathcal{M}_{NC} = \langle M, t_{ab}, h^{ab}, \nabla^{NC}, \rho \rangle$. Up to isomorphism, any two symmetry-related models of NGT correspond to a unique model of Newtonian gravity geometrised in this way. Thus, it is said, by moving to this "Newton-Cartan" theory one successfully removes the undesirable "gauge-redundancy" inherent in all non-geometrised versions of Newtonian gravitation theory.[18]

What might the defender of the interpretational view of symmetries say in defence of her view—in this context, that the move to Newton-Cartan theory is not required in order to be able to legitimately regard all symmetry-related solutions of NGT as physically equivalent?

I anticipate two likely lines of response. First, she might attempt to establish the preferability of her view over the motivational view by noting that the defender of the motivational view is committed, at least prior to the appropriate theory's reformulation (in the context of NGT), to the existence of in principle undetectable (symmetry-variant) matters of fact. Moreover, the defender of the interpretational view might argue, this is an unpalatable consequence, one which we would do best to avoid—and one which, she might point out, the interpretational view does in fact avoid.

I agree that the admission of such in principle undetectable facts is an undesirable consequence of the motivational view. However, I do not think that this admission is sufficiently unpalatable so as to be capable of refuting the motivational view, or even of establishing the preferability of the interpretational view over the motivational view. After all, prohibitively strong versions of verificationism aside, there is nothing obviously absurd about admitting in principle undetectable facts into one's ontology; nor is there any obvious reason why we should always be capable of discovering a theory, or a perspicuous characterisation thereof (the case of isomorphic models excepted), which succeeds in transparently explaining such solutions' empirical equivalence by virtue of

[18] For further details, see, e.g., Knox (2014). Note also the important point that moving to Newton-Cartan theory is not by itself sufficient for one to be able to transparently understand as physically equivalent all symmetry-related models of Newtonian theory set in flat spacetime. This is because—as mentioned above—such symmetry-related models will typically correspond to a single model of Newton-Cartan theory *only up to isomorphism*. Thus, in order to have a *fully* transparent understanding of how it is that symmetry-related models of Newtonian theory set in flat spacetime can correspond to a single model of Newton-Cartan theory, a modestly structuralist conception of spacetime ontology is also required.

their actual physical equivalence; nor indeed is there even any obvious way of guaranteeing that there will always be such a theory or characterisation (again, isomorphic models excepted) waiting in logical space to be discovered.

Furthermore, although it is to be admitted that the Newtonian who subscribes to the merely motivational view of symmetries might indeed be committed to the possibility of there being facts beyond her epistemic grasp, it nevertheless bears emphasising that for such a Newtonian there is a perfectly good explanation as to *why* such facts are epistemically inaccessible: they are inaccessible precisely because the world is in fact accurately described by the laws of NGT, with associated models of the form $\langle M, t_{ab}, h^{ab}, \sigma^a, \rho, \phi \rangle$, and because all any Newtonian observer ultimately has empirical access to are the relative distances and velocities between material entities. For such a Newtonian, then, the empirical phenomena underdetermine the genuine physical facts; but the theory itself is able to provide a perfectly transparent explanation of the reality behind the phenomena in terms of which the underdetermination can be straightforwardly understood.

The Newtonian who adopts the interpretational construal of symmetries, however, would appear to lose this explanatory transparency. In other words, she might know *that* she may legitimately regard all symmetry-related solutions as physically equivalent; but the reality in terms of which this physical equivalence is to be understood will (absent a reformulation of the theory) remain opaque to her; she is offered no immediate explanation as to *how* such physical equivalence is to be construed, or how it could even be said to arise.

These considerations naturally suggest a second possible line of response for the defender of the interpretational view. In particular, she might claim that she *does*, in fact, have a transparent understanding of the reality underlying NGT's symmetry-related models, and that such a transparent understanding is in fact attainable *prior* to the move to Newton-Cartan theory.[19]

Such a response evidently leads into deep philosophical waters very quickly. (After all, what does it mean to be in possession of a "transparent understanding" of anything?) But let me make a brief remark as to why I find this particular claim to be implausible. For note that in NGT the persisting points of absolute space are not merely "idly turning wheels" that can simply be expunged from the theory without explanatory loss: they are not "explanatorily idle" posits. This is for two main reasons. First, such points play a crucial role in the *metaphysical* explanation of what quantities like relative velocity and absolute rotation and absolute acceleration truly are: for the Newtonian, facts about particular inter-particle velocities and absolute rotations and absolute accelerations are naturally understood as being *grounded in* particular facts about (rates of change of) absolute velocities.[20] Second, such points provide the crucial transtemporal standard which is required in the realist's *causal* explanation of the observable effects of noninertial motion (e.g., Newton's famous "bucket experiment"): a standard without which Newton's laws simply cannot be formu-

[19] Dewar (2015, esp. 322)—who is a recent, explicit defender of the interpretational view—is plausibly read as making this claim.

[20] Cf. Pooley (MS, 118).

lated (at least, absent any *other* way of construing the transtemporal structure required to underwrite the distinction between inertial and noninertial motion). In short—and to the extent that the interpretational view is not supposed to reduce to a rather uninteresting form of scientific instrumentalism—it is simply not clear what causal-explanatory, *realistic* picture of the world is being propounded by the defender of the interpretational view, at least in this particular (Newtonian) context; it is simply opaque what, according to her, *the world is really like*.

Acknowledgements

For extremely helpful comments and discussion, I would like to thank Neil Dewar, James Ladyman, Niels Martens, Tushar Menon, Oliver Pooley, James Read, Simon Saunders, Alex Skinner, Teru Thomas, David Wallace, and audiences in London and Cardiff.

References

Dasgupta, S. (forthcoming), "Symmetry as an Epistemic Notion (Twice Over)." *British Journal for the Philosophy of Science.*

Dewar, N. (2015), "Symmetries and the Philosophy of Language." *Studies in the History and Philosophy of Modern Science*, Vol. 52, pp. 317-327.

Dirac, P. A. M. (1930), *The Principles of Quantum Mechanics*. Oxford University Press. (Reference is made to 1958 (4th) edition.)

Earman, J. (1989), *World-Enough and Space-Time*. MIT Press.

Greaves, H. and Wallace, D. (2014), "Empirical Consequences of Symmetries." *British Journal for the Philosophy of Science*, Vol. 65, No. 1, pp. 59-89.

Healey, R. (2007), *Gauging What's Real*. Oxford University Press.

Knox, E. (2014), "Newtonian Spacetime Structure In Light of the Equivalence Principle." *British Journal for the Philosophy of Science*, Vol. 65, No. 4, pp. 863-880.

Ladyman, J. (2007), "Scientific Structuralism: On the Identity and Diversity of Objects in a Structure." *Aristotelian Society Supplementary Volume*, Vol. 81, No. 1, pp. 23-43.

Maudlin, T. (1993), "Buckets of Water and Waves of Space: Why Spacetime Is Probably a Substance." *Philosophy of Science*, Vol. 68, No. 2, pp. 183-203.

Nozick, R. (2001), *Invariances: The Structure of the Objective World*. Harvard University Press.

O'Leary-Hawthorne, J. and Cortens, A. (1995), "Towards Ontological Nihilism." *Philosophical Studies*, Vol. 79, No. 2, pp. 143-165.

Pooley, O. (2013), "Substantivalist and Relationalist Approaches to Spacetime." In R. Batterman (ed.), *Oxford Handbook of Philosophy of Physics*. Oxford University Press.

Pooley, O. (MS), *The Reality of Spacetime*. Book manuscript.

Saunders, S. (2003), "Physics and Leibniz's Principles." In K. Brading & E. Castellani (eds.), *Symmetries in Physics: Philosophical Reflections.* Cambridge University Press.

Saunders, S. (2007), "Mirroring as an A Priori Symmetry." *Philosophy of Science*, Vol. 74, No. 4, pp. 452-480.

Weatherall, J. (forthcoming). "Understanding Gauge." *Philosophy of Science.*

Weyl, H. (1952), *Symmetry.* Princeton University Press.

Black Holes, Information Loss and the Measurement Problem

Elias Okon
Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México, Mexico City, Mexico.

Daniel Sudarsky
Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Mexico City, Mexico.

The *information loss paradox* is often presented as an unavoidable consequence of well-established physics. However, in order for a genuine paradox to ensue, not-trivial assumptions about, e.g., quantum effects on spacetime, are necessary. In this work we will be explicit about these additional, speculative assumptions required. We will also sketch a map of the available routes to tackle the issue, highlighting the, often overlooked, commitments demanded of each alternative. In particular, we will display the strong link between black holes, the issue of information loss and the measurement problem.

1 Introduction

The so-called *information loss paradox* is usually introduced as an unavoidable consequence of standard, well-established physics. The paradox is supposed to arise from a glaring conflict between Hawking's black hole radiation and the fact that time evolution in quantum mechanics preserves information. However, the truth is that, in order for a genuine paradox to appear, a sizable number of additional, non-standard assumptions is required. As we will see, these extra assumptions involve thesis regarding the fundamental nature of Hawking's radiation, guesses regarding quantum aspects of gravity and even considerations in the foundations of quantum theory.

In this work, we will be explicit about the additional assumptions required for a genuine conflict to arise and delineate the available options in order to tackle the issue. In particular, we will stress the connection between information loss and the measurement problem, and display the often non-trivial commitments that each of the available alternatives to solve the information loss issue demands.

2 The classical setting: black holes hide information

We start by reviewing some properties of classical black holes. Gravity, being always attractive, tends to draw matter together to form clusters. In fact, if the mass of a cluster is big enough, nothing will be able to stop the contraction until, eventually, a black hole will form. That is, the gravitational field at the surface of the body will be so strong that not even light will be able to escape and a region of spacetime from which nothing is able to emerge will form. The boundary of such a region is called the event horizon and, according to general relativity, its area never decreases.

In general, the collapse dynamics that leads to the formation of a black hole can, of course, be very complicated. However, it can be shown that all such systems eventually settle down into one of the few stationary black hole solutions, which are completely characterized by the mass, charge and angular momentum of the the Kerr-Newman spacetimes. In fact, the so-called black hole uniqueness theorems guarantee that, as long as one only considers gravitational and electromagnetic fields, then these solutions represent the complete class of stationary black holes. Moreover, the so-called no-hair theorems ensure that the set of stationary solutions does not grow, even if one considers other hypothetical fields.

The above mentioned results seem to suggest that when a cluster collapses to form a black hole, a large amount of information is lost. That is, details such as the multipole moments of the initial mass distribution, or the type of matter involved, seem to be altogether lost when the black hole settles. Note however that such apparent loss of information corresponds only to that available to observers outside of the black hole. While at early times there are Cauchy hypersurfaces[1] completely contained outside of the black hole, at later times all Cauchy hypersurfaces have parts both inside and outside it (see Figure 1). Therefore, using data located both outside and inside of the black hole, the *whole* spacetime can always be recovered. We conclude that, in the classical setting, information is not really lost. All that happens is that, when a black hole forms, a new region of no escape emerges and some of the information from the outside of the black hole moves into such new region. One could still argue that, since there are points inside of the horizon which are not in the past of future null infinity,[2]

[1] A Cauchy hypersurface is a subset of spacetime which is intersected exactly once by every inextensible, non-spacelike curve.

[2] Future null infinity is the set of points which are approached asymptotically by null rays which

then it is impossible to reconstruct the whole spacetime by evolving backwards the data on it. However, future null infinity is not a Cauchy hypersurface so one should not expect to reconstruct the whole spacetime from such data.

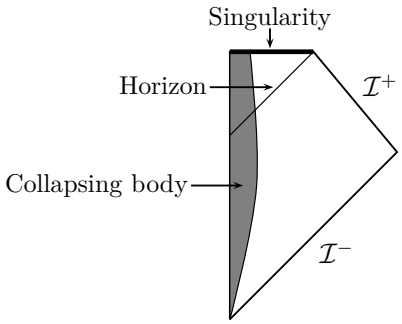


Figure 1: Penrose diagram for a collapsing spherical body. $\mathcal{I}^+$ and $\mathcal{I}^-$ denote past and future null infinity.

3 QFT on a fixed curved background: black holes radiate

The most dramatic change in our understanding of black hole physics came as a result of Hawking's famous analysis. What this analysis showed was that the formation of a black hole would modify the state of any quantum field in such a way that, at late times, there would be an outgoing flux of particles carrying energy towards infinity. Moreover, Hawking showed that the flux was characterized by the surface gravity κ of the resulting asymptotic stationary state of the black hole. This discovery transformed our perception of the formal analogy, originally pointed out in Bekenstein (1972), between the laws of black hole dynamics, and the standard laws of thermodynamics (see Wald (1994) for a discussion). In particular, it led to the view that the surface gravity is in fact a measure of the black hole's temperature $T = \frac{\kappa}{2\pi}$, and that the event horizon's area A is a measure of the black hole's entropy $S = A/4$.

Hawking's result is probably the most famous of the effects that arise from the natural extension of special relativistic quantum field theory to the realm of curved spacetimes. It imposes a dramatic modification on the classical view of black holes as

can escape to infinity.

absolutely black and eternal regions of spacetime. It is important to stress, though, that Hawking's calculation, being a result pertaining to quantum field theory on a *fixed* spacetime, does not encompass back-reaction effects. These are in fact notoriously difficult to deal with and a general framework for doing so is lacking. At any rate, some straightforward physical considerations, which have rather dramatic consequences, are often brought to bear in this context.

4 Back-reaction and first quantum gravity input: black holes evaporate

As can be expected, Hawking's result also suggests a dramatic modification in our expectation for the ultimate fate of a black hole. That is, while before Hawking's discovery, one would have expected that, once formed, a black hole would be eternal, the fact that the radiation is caring energy away, assuming overall energy conservation, leads one to expect that the mass of the black hole will start diminishing. The context in which this problem is standardly set is that of asymptotically flat spacetimes, for which we have a well defined notion of overall energy content given by the ADM mass[3] of the spacetime, a quantity which is known to be conserved.

As we noted, Hawking's calculation cannot deal with back-reaction. However, our confidence on energy conservation in the appropriate situations is so robust that it is difficult not to conclude that, as the radiation carries away energy, the black hole mass will have to diminishing. If this takes place, the surface gravity of the black hole—which is no longer really stationary, but can be expected to deviate from stationarity only to a very small degree—would change as well. As it turns out, the surface gravity is inversely proportional to the black hole's mass, so the black hole temperature can be expected to increase, leading to a ever more rapid rate of energy loss and a correspondingly faster decrease in mass.

The run away picture for the evaporation process suggests a complete disappearance of the black hole in a finite amount of time. Of course, we cannot really be sure about this picture because, in order to perform a solid analysis, we would need to deploy a, currently lacking, trustworthy theoretical formalism adept to the challenge. The

[3]The ADM mass is a quantity associated with the asymptotic behavior of the induced spatial metric of a Cauchy hypersurface. In asymptotically flat spacetimes, it is known to be independent of the hypersurface on which it is evaluated (see Arnowitt et al. (1962)).

problem is that, by the removal of energy from the black hole, one can expect to eventually reach a regime where quantum aspects of gravitation become essential to the description of the process. At such point, one might contemplate the possibility that, as a result of purely quantum gravitational aspects, the Hawking evaporation of the black hole will stop, leaving a small stable remnant. This, in turn, might open certain possibilities regarding the information issue. For the time being, though, we will ignore such an option.

Then, in order to simplify the discussion at this point, we will ignore the possibility of remnants and assume that there is nothing to stop the Hawking radiation. Then, if the black hole's mass decreases in accordance with energy conservation, one expects that the black hole to simply disappear and the spacetime region where it was located to turn flat (see Figure 2).

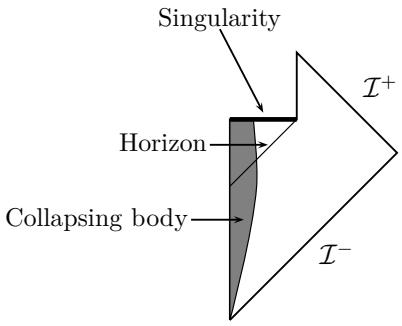


Figure 2: Penrose diagram for a collapsing spherical body, taking into account Hawking's radiation.

At this point, we seem to come face to face with an information loss problem: the original massive object that collapses, leading to the formation of a black hole, might have required an incredibly large amount of detail for its description. However, the final state that results from the evaporation is simply described in terms of the thermal Hawking flux, followed by an empty region of spacetime. More to the point, even if the initial matter that collapses to form a black hole was initially in a pure quantum state, after the complete evaporation of the black hole there would be a mixed one, corresponding to the thermal Hawking flux. These considerations seem to indicate that, even at the fundamental level, we have a fundamental loss of information. The final state, even if described in full detail, does not encode the information required to retrodict the details of the initial one. At the level of quantum theory, we would

be facing a non-unitary (and non-deterministic) relation between the initial and final states of the system, a situation that seems at odds with the unitary evolution provided by the Schrödinger equation.

There are, however, various caveats to the above conclusion. The first one is opened up by the possibility of the evaporation eventually stopping, leading to a stable remnant. The mass of said remnant can be estimated by considering the natural scales at which the effects of quantum gravity are expected to become important. This leads to an estimate of the order of Plank's mass ($\approx 10^{-5}$ gr). Then, if one wants the remnant to encode all the information present in the initial state, one is led to the conclusion that such a small object would have a number of possible internal states as large as that of the original matter that collapsed to form the black hole, which can, of course, have had a mass as large as one can imagine. It is hard, then, to envisage what kind of object, with such rather unusual thermodynamical behavior, would this remnant have to be. For this reason, this possibility is usually not considered viable (although we acknowledge that these considerations might be overturned; for a discussion of these issues see Banks (1994)). At any rate, we will not consider this possibility any further.

We should also mention another proposal which uses the idea that, while curing singularities, quantum gravity might open paths to other universes, which could be home to the missing information. Such information would be encoded either in a new universe or in correlations between it and ours. Besides the dramatic ontological burden, such proposal leaves open the possibility of these alternative universes emerging even in ordinary processes (which could, e.g., involve virtual black holes), leading to information loss in such standard scenarios. Alternatively, the information could be preserved, but impossible to retrieve in principle. We will also not consider this possibility any further.

A much more important caveat is the following: we have very solid results indicating that, associated with the formation of a black hole, there is always a singularity of spacetime appearing withing it. The strongest results in this regard are a series of theorems proved by Hawking (see Hawking and Ellis (1973)) showing that, under quite general conditions, and assuming reasonable properties for the energy and momentum of the collapsing matter, the formation of singularities is an inevitable result of Einstein's equations. The issue is that, at the classical level, these singularities represent a breakdown of the theory and, in fact, a failure of the spacetime description. The singularities are, therefore, to be thought of as representing boundaries of spacetime, rather that points within it. Once a spacetime has additional boundaries, it is clear

that the issue of information has to be confronted on a different light. Of course, if one considers the description of the system at an initial Cauchy hypersurface and wants a final hypersurface to encode the same information, one has to make sure that the final one is also Cauchy.

The formation of singularities then implies that, if we want to have spacetime regions where the system's state could be thought of as encoding all the information, then we must surround the singularities by suitable boundaries. In other words, if the singularities force us to include further boundaries of spacetime, then the comparison of initial and final information has to be done between the initial Cauchy hypersurface and the late-time *collection* of surfaces that, together, act as a Cauchy hypersurface. That collection could naturally include asymptotically null future, but also the hypersurfaces surrounding the singularities. The same kind of calculation as the one done by Hawking would then show that all the information present on the initial hypersurface would also be encoded in the state associated with this late-time Cauchy hypersurface. That is, if we include the boundary of spacetime that arises in association with the singularity, then there is no issue regarding the fate of information. We conclude that, under these circumstances, still there is no information loss.

5 Second quantum gravity input: black holes do not involve singularities

As we noted above, singularities represent a breakdown of the spacetime description as provided by general relativity and thus indicate the need to go beyond such theory. The expectation among theorists is that quantum gravity is going to be the theory that cures these failures of classical general relativity, replacing the singularities by a description in the language appropriate to quantum gravity. This is, in fact, what occurs with various other theories that are known to be just effective descriptions of a physical system's behavior in a limited context, but that have to be replaced with a more fundamental description once the system leaves that regime. Think for instance of the description of a fluid by, say, the Navier-Stokes equations. We know that this description works very well in a large variety of circumstances, but that a breakdown of such description occurs, for instance, when there are shock waves or when other types of singularities are formed. However, under such circumstances, the underlying kinetic theory, including the complex inter-molecular forces, is expected to remain valid. The

point is that, just as in those cases, one expects the emergence of singularities in general relativity to indicate the end of the regime where the classical description of spacetime is valid and, therefore, where a quantum gravity description would have to take over (see Figure 3 and Ashtekar and Bojowald (2005) for details).

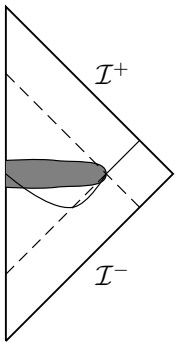


Figure 3: "Quantum spacetime diagram" for a black hole.

Of course, if quantum gravity does in fact cure the singularities, and removes the need to consider, in association with the corresponding regions, a boundary of spacetime, the issue of the fate of information in the Hawking evaporation of black holes resurfaces with dramatic force. So, do we finally have a genuine paradox in our hands. Not quite yet; a few elements are still missing. In order for a paradox to arise, we need to couple a genuine loss of information with a fundamental theory which does not allow for information to be lost.

6 A paradox?

When is it, then, that the Hawking radiation by a black hole leads to an actual paradox? We are finally in a position to enumerate the various assumptions required in order to construct a genuine conflict:

1. As a result of Hawking's radiation carrying energy away from the black hole, the mass of the black hole decreases and it either evaporates completely or leaves a small remnant.

2. In the case where the black hole leaves a small remnant, the number of its internal degrees of freedom is bounded by its mass in such a way that these cannot possibly encode the information contained in an arbitrarily massive initial state.

3. Information is not transfered to a parallel universe.

4. As a result of quantum gravity effects, the internal singularities within black holes are cured and replaced by something that eliminates the need to consider internal boundaries of spacetime.

5. The outgoing radiation does not encode the initial information.

6. Quantum evolution is always unitary.

We have already discussed the arguments in support of assumptions 1, 2, 3 and 4 and saw that, although by no means conclusive, they are reasonable. But what about 5 and 6? Well, in order to avoid a paradox, and assuming the first four assumptions to be true, at least one of them has to be negated. In order to explore the motivations and consequences of doing so, we must think clearly about how to interpret Hawking's calculation in a context in which 1, 2, 3 and 4 are the case.

As we remarked above, Hawking's calculation is performed in the setting of a quantum field theory over a fixed curved background. What one finds there is that an initial pure state of the field evolves into a final one which, when tracing over the inside region, reduces to a mixed thermal state. The key question at this point, then, is how to interpret such a final mixed state in a setting in which i) the black hole is no longer there, so there is no interior region to trace over, and ii) in which there is no singularity (or parallel universe) for the information to "escape into." As far as we can see, there are two alternatives: either one assumes that the mixed state arises only as a result of tracing over the interior region and maintains that the outgoing radiation somehow encodes the initial information—which amounts to negating 5; or one takes Hawking's result seriously and maintains that, even in this scenario, information is lost—which amounts to negating 6. Below we explore each option in detail.

6.1 The outgoing radiation encodes information

In the last couple of decades, the community's position on the information loss subject has been strongly influenced by developments in String theory. Such framework has permitted exploration of questions, regarding black holes, using settings where event horizons and singularities play no relevant roles. This is possible due to the AdS/CFT correspondence (see e.g., Strominger (2001)), which allows the mapping of complicated spacetime geometries in the "bulk" of asymptotically Anti-de Sitter spacetimes,

including ones involving black holes, onto corresponding states of an ordinary quantum field theory living on the Anti-de Sitter boundary (which is, in fact, a flat spacetime). These considerations have led people to conclude that, as a breakdown of unitarity is not expected to take place in the context of a quantum field theory in flat spacetimes, there should be no room for a breakdown of unitarity in the corresponding situation involving black holes either.[4]

The proposal, then, is that unitarity is never broken and that information is never lost. As a result, Hawking's calculation has to be somehow attuned to assure consistency. In particular, the proposal is that the outgoing radiation must encoded all of the initial information. There is, however, a high price to pay in order to achieve this. As has been shown in Almheiri et al. (2013), in order for the outgoing radiation to encode the necessary information, each emitted particle must get entangled with all the radiation emitted before it. However, due to the so-called, "monogamy of entanglement," doing so entails the release of an enormous amount of energy, turning the event horizon into a *firewall* that burns anything falling through it. The upshot then, is a divergence of the energy-momentum tensor of the field over the event horizon and a radical breakdown of the equivalence principle over such a region.

6.2 Unitarity is broken

The discovery of the Hawking radiation was initially taken as a clear indicative of information loss at the fundamental level. In fact, Hawking (1976) even introduced a notation for this general type of evolution which was supposed to account for the transformation from (possibly pure) initial states ρ_i into final mixed ones ρ_f. Hawking denoted the general linear, non-unitary, operator characterizing such transformation by the sign \$, i.e., $\rho_f = \$\rho_i$. Likewise, Penrose pointed out that, in order to have a consistent picture of phase space for situations involving black holes in thermal equilibrium with an environment, one has to assume that ordinary quantum systems undergo something akin to a self-measurement, by which he meant quantum state reduction that was not the result of measurement by external observers or measuring devices (see Penrose (1981)). Penrose (1999) further argued that quantum state reduction is probably linked to aspects of quantum gravity.

The early assessments of these ideas in Banks et al. (1984) indicated that they

[4]Note however that the argument can be easily reversed to show exactly the opposite. Since Hawking's result shows that unitarity breaks when black holes are present, one must conclude that quantum evolution *cannot* be unitary even in a quantum field theory on flat spacetimes.

where likely to lead to a very serious conflict with energy and momentum conservation or to generate unacceptable non-local features in ordinary physical situations. However, further analysis in Unruh and Wald (1995) showed that these assessments where not that solid and that there where various possibilities to evade the apparently damning conclusions.

In (omitted references) we have explored the viability of breaking unitarity both qualitatively and quantitatively. In particular, we have successfully adapted objective collapse models, developed in connection with foundational issues within quantum theory, in order to explicitly describe the transition from the initial pure state into a mixed one. Our view on the subject is based on the conviction that, contrary to the prevailing opinion in the community working on the gravity/quantum interface, there are good reasons to think that quantum theory requires modifications to deal with its basic conceptual difficulties. Below we discuss these issues and explore their consequences for the information loss paradox.

7 Information loss and the measurement problem

Most discussions of black holes and information loss do not implicate foundational issues of quantum theory. Of course, ignoring such issues, particularly with pragmatic interests in mind, is often acceptable. However, when deep conceptual questions are involved, such as in the present case, the pragmatic attitude might not be the right way to go.

The standard interpretation of quantum mechanics involves a profoundly *instrumentalist* character, with notions such as *observer* or *measurement* playing a crucial role. Such an instrumentalist trait becomes a problem as soon as one intends to regard the theory as a fundamental one, useful not only to make predictions in suitable experimental settings, but also to be applied to the measurement apparatuses, to the observers involved, or to non-standard contexts such as black holes or the universe as a whole. The resulting problem, often referred to as the *measurement problem*, has been discussed at length in numerous places and many different concrete formulations of it have been given. A particularly useful way to state it, given in Maudlin (1995), is as a list of three statements that cannot be all true at the same time:

A. The physical description given by the quantum state is complete.

B. Quantum evolution is always unitary.

C. Measurements always yield definite results.

Maudlin's formulation of the measurement problem is noteworthy because of its generality and its preciseness. Moreover, it is extremely useful in order to motivate and classify strategies to solve the problem. For example, by negating A, one arrives at so-called hidden variable theories, such as Bohmian mechanics; by removing B, one gets so-called objective collapse theories, such as GRW; and by discarding C, Everettian interpretations emerge. Of these three options, the last one is, by far, the most contentious. Among its most urgent matters, we can mention the problem of the preferred basis, the one of making sense of probabilities in the theory and the general and basic issue of establishing a clear and precise link between the abstract mathematical objects of the theory and concrete empirical predictions. Of course, brave attempts to deal with these and other issues within Everettian frameworks abound. However, be believe that, at least for the time being, they are far from being successful.

Returning to the measurement problem and its relation to the information loss issue, we note that assumptions 6 and B are in fact identical. Therefore, the strategy one decides to adopt in order to avoid complications regarding the information loss issue (e.g., negating 5 or 6 above) has implications with respect to what one must say regarding the measurement problem (e.g., negating A, B or C). In particular, if regarding the information loss, one decides to maintain the validity of 6 (and thus to hold that the outgoing radiation encodes all of the initial information), then one necessarily has to either negate A or C (i.e., either to entertain a hidden variables theory or an Everettian scenario). In other words, insisting on a purely unitary evolution, not only demands a violation of the equivalence principle and a divergence of the energy-momentum tensor, but also a commitment either with many worlds or with an acknowledgment that standard quantum mechanics is incomplete. On the other hand, if regarding the information loss problem, one decides to abandon unitarity, the same move automatically not only avoids a breakdown of the equivalence principle, but also guarantees success with respect to the measurement problem. The upper hand of the second option seems evident to us.

8 Conclusions

Since the publication of Hawking's analysis, more than forty years ago, the issue of black hole information loss has been a central topic in theoretical physics. The AdS/CFT

correspondence, proposed almost twenty years latter, came to further propel an already notorious debate. Yet, even after all these years, the discussion is often engulfed by confusion and misunderstanding among participants. The objective of this work is to develop a clear analysis of some of the key conceptual issues involved. Our hope is that, by doing so, significant progress on this important topic could soon be achieved.

We have presented the basic theoretical setting of the black hole information issue, paying special attention to elements, arising from not yet well-established physics, that presently have to be regarded merely as reasonable assumptions. Moreover, we have argued that the information loss issue is closely related to the measurement problem, and claimed that it is precisely within the context of certain proposals put forward to deal with the latter that the former finds one of its most conservative resolutions.

References

Almheiri, A., Marolf, D., Polchinski, J., and Sully, J. (2013). Black holes: complementarity or firewalls? *JHEP*, 62.

Arnowitt, R., Deser, S., and Misner, C. (1962). The dynamics of general relativity. In Witten, L., editor, *Gravitation: an introduction to current research*. Wiley.

Ashtekar, A. and Bojowald, M. (2005). Black hole evaporation: a paradigm. *Class. Quant. Grav.*, 22(3349).

Banks, T. (1994). Lectures on black hole information loss. *Nucl. Phys. Proc.*, 41.

Banks, T., Susskind, L., and Preskin, M. E. (1984). Difficulties for the evolution of pure states unto mixed states. *Nucl. Phys. B*, 224(125).

Bekenstein, J. D. (1972). Black holes and the second law. *Lett. Nuovo Cim.*, 4(737).

Hawking, S. W. (1976). Breakdown of predictability in gravitational collapse. *Phys. Rev. D*, 14(2460).

Hawking, S. W. and Ellis, G. F. R. (1973). *The large scale structure of spacetime*. Cambridge University Press.

Maudlin, T. (1995). Three measurement problems. *Topoi*, 14.

Penrose, R. (1981). Time asymmetry and quantum gravity. In Isham, C. J., Penrose, R., and Sciama, D. W., editors, *Quantum Gravity II*. Clarendon Press.

Penrose, R. (1999). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press.

Strominger, A. (2001). The AdS/CFT correspondence. *JHEP*, 0110(034).

Unruh, W. G. and Wald, R. M. (1995). On evolution laws taking pure states to mixed states in quantum field theory. *Phys. Rev. D*, 52:2176–2182.

Wald, R. M. (1994). *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics*. University of Chicago Press.

The Causal Homology Concept

Jun Otsuka*

Abstract

This presentation proposes a new account of homology, which defines homology as a correspondence of developmental or behavioral mechanisms due to common ancestry. The idea is formally presented as isomorphism of causal graphs over lineages. The formal treatment not only clears the metaphysical skepticism regarding the homology thinking, but also provides a theoretical underpinning to the concepts like constraints, evolvability, and novelty. The novel interpretation of homology suggests a general perspective that accommodates evolutionary developmental biology (Evo-Devo) and traditional population genetics as distinct but complementary approaches to understand evolution, facilitating further empirical and theoretical researches.

*Department of Philosophy, Kobe University, Rokko-dai 1-1 Nada, Kobe, Japan. Email: junotk@gmail.com

1 Introduction

The homology thinking, the idea that the same anatomical structure repeatedly appears in different species or parts of the same organism, has a long history in biology (Amundson, 2005). While the existence of such anatomical similarities among or within species is now explained by the descent from a common ancestor, the conceptual issues surrounding the notion have invited philosophical as well as methodological debates and skepticism. Owen famously defined homology as "the same organ in different animals under every variety of form and function," but this definition is perplexing rather than enlightening: what characterizes and warrants the sameness of "organs," if not their form or function? What, in other words, is the unit of homology?

There are three conceptual problems. The first and foremost problem is its *definition*: what exactly is homology? Evolutionary theory tells us that homology is identity due to a common origin, but an identity of *what*? Is it morphological characters, activities, clusters of properties, or genetic networks that are regarded to be same? And what is the criterion to judge whether or not two such things are actually the "same"? The second problem is *metaphysical*. As Ghiselin (1997) points out, the homology-as-identity partitions the whole tree of life into equivalence classes. But doesn't the supposition of such universal classes, reminiscent of Aristotelian essence, commit us to an anti-evolutionary thinking? And thirdly, there is a *pragmatic* question: why do we care about homology at all? Some neo-Darwinians such

as G. C. Williams see homologs as mere "residues," i.e. a relic of the past common ancestry not yet washed out by natural selection (Amundson, 2005, pp. 237-8). If that is the case homology by itself would have no explanatory role in evolutionary theory, and the quest for its definition, however well-defined and metaphysically sound, becomes a mere armchair exercise with no scientific value.

There is at least one usage of the concept free from these issues: homology of DNA sequences. Here the "sameness" is well-defined by matching bases that can be one of the four chemical kinds, G, C, T, A. Moreover, the scientific importance of orthologs and paralogs is undeniable in reconstructing the evolutionary history and predicting gene function, to name a few. Things become different for phenotype, in particular complex phenotypes like morphological or behavioral traits. First of all, there is no clear-cut definition of "phenotypic units" as that for nucleotides. Continuous traits such as height or weight usually lack objects breakpoints by which we classify them into discrete equivalence classes. In sum, there seem to be no non-arbitrary and non-controversial units for phenotype of which we can talk about the sameness, and thus homology.

Our first task, therefore, is to identify the units on which the phenotypic homology relationship can be defined. This presentation proposes that this purpose is best served by *causal graphs* which formally represent developmental or behavioral mechanisms. Homology is thus defined as graph isomorphism over lineages, or conservation of the underlying causal structure

over evolutionary history (Section 2). I will argue in Section 3 that the formal treatment of homology (i) solves the philosophical as well as empirical puzzles and criticisms regarding the homology concept; (ii) provides clear meanings to some key but elusive concepts such as constraints, evolvability, and novelty; (iii) and suggests a broad perspective that accommodates evolutionary developmental biology (Evo-Devo) and traditional population genetics as distinct but complementary research projects. Section 4 compares the present approach to other existing accounts of homology, and discusses its relative strengths, challenge, and philosophical implication. As will be stressed there, the primary objective of this presentation is to facilitate or open up new empirical as well as theoretical questions. The last section concludes with some of these research prospects that are prompted by the new homology concept.

2 Defining homology with graphs

The idea of characterizing homology in terms of causal structures is not new. Various biologists have suggested, albeit in different fashions, that the developmental or behavioral mechanisms underlying phenotype can or should serve as a unit of homology (e.g. Riedl, 1978; Wagner, 1989, 2014; Gilbert and Bolker, 2001; Müller, 2003). These proposal, however, are mostly based on independent examples or qualitative descriptions, and the lack of a unified treatment has blurred their philosophical as well as theoretical implications.

The aim of this section is to give a formal representation to the ideas of developmental sameness by using causal graphs, in view of exploring the conceptual nature of homology in the later sections.

A *causal graph* $\mathcal{G}$ is a pair $(\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a set of phenotypic or genetic variables of organisms and $\mathbf{E}$ is a set of edges representing causal relationships among these traits. Development is understood as a causal web connecting embryological, morphological, and behavioral traits, and the set of edges $\mathbf{E}$ characterizes these causal links. Note that such connections may remain invariant even under considerable modifications in phenotypic values or the functional form that determines the quantitative nature of each edge. The same set of $\mathbf{E}$ is consistent with a variety of phenotypic states and forms of causal production; it only defines the qualitative feature of the causal networks, i.e. which causes which.

Once modeled in this way, it becomes meaningful to compare causal structures of different organisms. A causal graph $\mathcal{G}_1 = (\mathbf{V}_1, \mathbf{E}_1)$ is *isomorphic* to another $\mathcal{G}_2 = (\mathbf{V}_2, \mathbf{E}_2)$ if they have the same structure, or more formally if there is a bijection $f : \mathbf{V}_1 \to \mathbf{V}_2$ such that if $(v, w) \in \mathbf{E}_1$ then $(f(v), f(w)) \in \mathbf{E}_2$. Likewise, isomorphism can be defined for subgraphs, which are just parts of the causal graphs restricted to a subset $\mathbf{V}' \subset \mathbf{V}$. We write $\mathcal{G}_1 \sim \mathcal{G}_2$ if two (sub)graphs are isomorphic. It is easy to see '$\sim$' is symmetric, reflexive, and transitive, and thus defines a equivalence class.

Each individual is assigned one causal graph that models a particular part of its developmental or behavioral mechanism. Let us denote the causal

structure of an organism a by $\mathcal{G}(a)$. Collectively, $\mathcal{G}(A)$ is a set of causal structures for a set of organisms A. We assume usual ancestor/descendant relationships over a set of organism Ω (which may include more than one species). If b is an ancestor of a, the *lineage* between b and a is a set of every individual between them. Given this setup homology is defined as follows.

> For two sets of organisms $A, B \subset \Omega$, let $\mathcal{G}'$ be a subgraph of all $g \in \mathcal{G}(A)$, and $\mathcal{G}''$ be a subgraph of all $g \in \mathcal{G}(B)$. Then $\mathcal{G}'$ and $\mathcal{G}''$ are homologous iff
>
> 1. $\mathcal{G}' \sim \mathcal{G}''$;
> 2. there is a set of common ancestors $C \subset \Omega$ of A and B[1]; and
> 3. for every d in all the lineages from C to A and C to B, $\mathcal{G}(d)$ has a subgraph $\mathcal{G}'''$ such that $\mathcal{G}''' \sim \mathcal{G}' \sim \mathcal{G}''$.

The definition explicates the idea that homology is the identity between causal structures due to common ancestry. Two (sets of) organisms share a homologous causal structure if, in addition to the graph isomorphism, every individual on the lineage connecting them shares the same causal graph, capturing the idea that the structure has been conserved through the evolutionary history.

The same treatment applies to serial homology, i.e. the homology relationship among parts of the same organism, such as teeth, limbs, or tree

[1]Note that C may be A or B themselves. Also note the condition 1 is redundant if a lineage includes the both ends. But here it is retained for clarity.

leaves. We can just set $A = B$, and compare different but isomorphic subparts $\mathcal{G}', \mathcal{G}''$ of the same overall structure $\mathcal{G}(A)$. Then the homology hypothesis is that there is an organism c in which the mechanism in question was duplicated, and the lineages from c to A have conserved the duplicated structures.

The above definition is illustrated with a case of special homology in figure 1, which depicts a particular region of the tree of life for (groups of) organisms A to G. Two mutations M_1, M_2 on the developmental mechanism occurred in the lineage leading to F, in which one causal edge $V_1 \to V_3$ was first removed and then restored. In this example, the causal structure $\mathcal{G}(D)$ of population D is homologous to $\mathcal{G}(E)$, for they are both inherited from the ancestral graph $\mathcal{G}(B)$ and $\mathcal{G}(A)$. In contrast, it is not homologous to $\mathcal{G}(F)$ even though they are graph-isomorphic. This is because the lineages connecting D and F do not conserve the causal structure in question: particularly it is not shared by C.

The example, though too simplistic to capture any real biological phenomena, makes explicit the idea that homology is a concordance of developmental mechanisms due to common ancestry. Note the criterion makes no reference to the resulting phenotype represented by particular values or distributions of variables. It does not require or forbid that, for example, two populations E and D show similar morphological distributions. Nor does it assume the graphs consist of the variables of the same nature. If the causal graphs in figure 1 represent a genetic network, kinds of genes/variables that

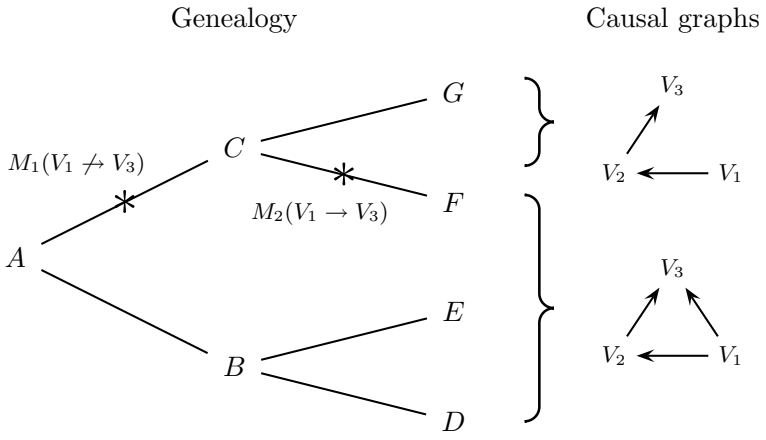


Figure 1: Illustration of graph homology. On the left is a genealogy tree for hypothetical populations A, B, C, D, E, F, G, while the graphs on the right describe causal structures of these populations over three characters, V_1, V_2, and V_3. Two asterisks ($*$) on the tree denote mutation events on the causal structure. See text for explanation.

constitute the network may vary across populations, as long as they serve the same causal roles within the overall structure. It is structural, rather than material, identity that defines homology. Theoretical as well as philosophical implications of this view will be explored in the following sections.

3 Conceptual advantages of the view

The above account is intended to provide a theoretical platform to formulate and evaluate hypotheses or explanations regarding homology. This section explicates the conceptual benefits of thinking homology in terms of causal graphs. Discussions on the empirical adequacy are differed to the next sec-

tion.

As discussed in the introduction, the major obstacle in defining homology is the absence of definite phenotypic units. Homology is an identity rather than similarity relationship (e.g. Ghiselin, 1997; Müller, 2003; Wagner, 2014), whereas no two or more phenotypic characters are identical in a strict sense — there are always subtle differences in, say, shape or size. The problem could be solved if we could find a natural and non-arbitrary way to factorize the phenotypic space into discrete regions so that two phenotypes within the same region are regarded "identical" despite their apparent differences. This is a difficult task, especially because we do not know the topological feature of the phenotypic space (Wagner and Stadler, 2003). To solve this issue the present analysis adopts a different strategy: instead of trying to impose a certain structure on the phenotypic space, it takes the generative mechanisms as basic units. Once these mechanisms are represented by causal graphs, which by nature are discrete mathematical entities, the desired identity relationship is given by graph isomorphism regardless of differences in the resulting morphology/phenotype. The graphical representation thus provides natural units prerequisite to define homology.

It is granted that a graph representation is not determined uniquely, because the same developmental mechanism can be modeled in various levels of abstraction, yielding causal graphs of different complexities. However, I take this to be a strength rather than weakness of my view, because homology too is often treated as description-dependent. Teleost fins and tetrapod limbs

are said to be homologous *as* paired vertebrate appendages, but *not as* fins or limbs. In contrast, our hands and pectoral fins of the whale are homologous not only as appendages but also as limbs. One tempting hypothesis is that such degrees of homology relationship correspond to isomorphisms of causal structures described at different granularities. In the above example, it is hypothesized that teleost fins and tetrapod limbs are represented by the same, but rather course-grained, causal graph, while tetrapod species share the causal structure to much finer details.

Fixing the level of abstraction determines not only the equivalent classes but also the degree of similarity between these classes. Two distinct causal graphs may be closer or further depending on the number of changes required to obtain one from the other. If $\mathcal{G}''$ is obtained by removing one edge from $\mathcal{G}'$ which in turn lacks one of the edges of $\mathcal{G}$, $\mathcal{G}''$ is one step further than $\mathcal{G}'$ from the original $\mathcal{G}$. Each such deletion or addition of causal connection is called *novelty*. Novelty in this framework is a modification of the causal graph, and as such creates a new equivalence class of causal graphs, namely homology. Evolutionary novelty also comes in different degrees. In general, a single modification in abstract graphs will correspond to multiple edge additions or deletions in detailed ones, and thus is weighted more. In this regard a change in the causal graph shared both by teleosts and tetrapods will count as a significant novelty and possibly a creation of a new "bauplan."

This brings us to one of the central contentions in today's evolutionary biology, namely the alleged inadequacy of the Modern Synthesis framework,

in particular population genetics, to incorporate macro-scale evolutionary phenomena uncovered by evolutionary developmental biology (e.g. Pigliucci and Müller, 2010). It has been claimed that homology (macro-scale conservatism) and novelty (a large phenotypic change) not only resist explanations by the Neo-Darwinian gradualism, but also constrain evolutionary trajectories as modeled in population genetics (e.g Amundson, 2005; Brigandt, 2007). The theoretical relationship between Evo-Devo and population genetics, however, remains elusive, which makes difficult to evaluate the call for the "new synthesis."

The present approach, by expressing homolgy and novelty in terms of graph equivalence and modification, suggests a perspective on this connection and a way to turn these claims into empirical hypotheses. Because causal models induce evolutionary changes as studied in population and quantitative genetics (Otsuka, 2015, 2016), the graphical representation allows one to analyze how developmental structures generate and constrain evolutionary dynamics. In particular, topological features of the graph such as modularity yield, via the so-called Markov condition, patterns of probabilistic independence on the phenotypic distribution and determine possible evolutionary trajectories or *evolvability*. The causal graph approach thus supports the view that a homolog constitutes a unit of morphological evolvability (Brigandt, 2007).

The graph structures that yield population dynamics are usually not study objects of population genetics. They rather serve as background frame-

works in which evolutionary models are build to study changes in genetic or phenotypic frequencies. These frameworks, however, must come from somewhere, and this evolutionary process is a primary interest of Evo-Devo. Studies on homology and novelty — graph stasis and change — amount to "higher order" evolutionary analyses that deal with changes in the theoretical framework used in population genetics to predict local population dynamics. The graphical conception of homology thus suggests a broad perspective that accommodates these different, and sometimes seen antagonistic, research fields as complementary approaches to understand evolution.

Finally, let us turn to the metaphysical problem. As seen above, homology is defined as an equivalence class over a set of causal graphs. But to what do such classes correspond, if not some ideal types or essences? Homology thinking has been criticized as anti-evolutionary due to its alleged commitment to essentialism. These critics thus re-interpret homology as a lineage that connects individual parts, rather than as a universal class to be instantiated by its members/homologs (e.g. Ghiselin, 1997). A detailed examination of this criticism must await another occasion, but here I just want to propose a different way to look at the issue. A metaphysical implication from the present study is that homology stands to concrete parts of organisms not as a universal to individuals, nor as a whole to parts, but rather as a model to phenomena to be modeled. A homology hypothesis is based on an observation that two or more individuals or parts thereof can be modeled by

the same causal graph.[2] Hence the proper relationship is not instantiation or mereology, but representation (Suppes, 2002). Once conceived in this way, the metaphysical ghost of essentialism vanishes away. Just like the same oscillator model characterizes various kinds of pendulum clocks, homology-as-model is a mathematical entity (directed graph) that may represent more than one actual individual, but that does not force us to commit to any form of essentialism.

The individual-universal distinction has also cast a shadow on the pragmatic issue regarding the epistemic role and significance of the concept of homology. It has been argued that the study of homology cannot be any more than a historiography since there is no such thing as a law for individuals (Ghiselin, 1997). A very different picture, however, emerges from the present thesis. A homology statement is a historical hypothesis regarding causal isomorphism — that two or more (sets of) organismal parts can be represented by the same causal model — and as such makes various predictions. For example, it supports extrapolations from model organisms, predicting that homologous organs will respond in the same or similar fashion to physiological, chemical, or genetic interventions. In addition, since isomorphic developmental structures will generate similar patterns of phenotypic variation (see above), their evolutionary changes are expected to follow similar trajectories. Establishing homologous relationships therefore is not a mere

[2]This, in turn, implies these individuals would respond in a more or less same fashion to hypothetical interventions (Woodward, 2003). Hence homology statements eventually boil down to counterfactual claims.

historical description, but has predictive implications both on physiological and evolutionary studies.

4 Comparisons and possible objections

This section compares the present proposal with some of the existing accounts of homology and also discusses possible objections. A number of philosophers and biologists have recently proposed to define homology as a *homeostatic property cluster*, a cluster of correlated properties maintained by "homeostatic mechanisms" (e.g. Boyd, 1991; Rieppel, 2005; Brigandt, 2009; Love, 2009). Since clustering and correlations are a matter of degree, homology according to this view is not an identity but a similarity relationship. It thus confronts with the boundary problem — to what extent properties must be clustered to form a homolog? The underlying "homeostatic mechanism" is supposed to clarify this boundary, but without a clear definition of what it is such an attempt only leads to a circularity. In particular, if it is defined as "those causal processes that determine the boundary and integrity of the kind (Brigandt, 2009, p.82)," the charge of circularity cannot be avoided.

This kind of problem will not arise if the generative mechanisms are defined explicitly in terms of causal graphs. While my approach proposes a formal framework to represent these mechanisms, it does not make any assumption or restriction on their structure: in particular it does not require the mechanism to be homeostatic, circumventing the criticism that a home-

ostatic mechanism by definition cannot evolve (Kluge, 2003). Moreover, the reference to "clusters" or even properties becomes superfluous, because the variational properties of phenotype are mere derivatives of the underlying causal graph. Of course, covarying traits suggest some ontogenetic connections, and thus may serve as a useful heuristics for finding homologs. They are, however, only "symptoms" — what *define* homology are not properties, clustered or homeostatic, but rather generative mechanisms.

The present approach has a closer affinity to the so-called *biological homology concept* that attempts to explain the phenomena of homology on the basis of a particular feature of the underlying causal structure, such as gene regulatory networks (e.g. Wagner, 1989, 2014). Indeed, one motivation of this presentation is to give a formal platform for these empirical hypotheses to elucidate their theoretical as well as philosophical implications. An important empirical challenge to the biological homology concept, and any other attempts to identify a homolog with a certain developmental structure, is the well-known fact that morphological similarity does not entail developmental sameness (Wagner and Misof, 1993). It has been reported that apparently homologous characters in related species may develop from different genes, cell populations, or pathways — the phenomena called *developmental system drift* (True and Haag, 2001). Although these phenomena present a challenge to my account as well, not all of them count as counter evidence. If, for example, "drift" concerns only genetic or cell materials, topological features of the causal network may remain invariant. Descriptive levels also matter.

Even if two causal structures differ at a fine-grained description, they may coincide at a more abstract level. Finally, my view does not require the entire developmental system to be conserved: if causal graphs share *some* part, they may still be homologous *in that aspect*. Indeed, it would be surprising if two apparent homologs turn out to share no developmental underpinnings at all. Some degree of flexibility may be expected, but so is inflexibility. Representing and comparing homologs in terms of the underlying causal graphs will serve as a heuristics to identify which part of the overall developmental system is responsible for generating similar morphological patterns.

From a philosophical perspective, a distinguishing feature of my account is its explicit reference to *models*. Homology has traditionally considered to be a relationship among concrete biological entities or properties thereof: it is organs or phenotypic features that are said to be homologous. In contrast, homology in my view is a relationship among abstract entities, i.e. causal graphs. How and why does such an abstract relationship reveal anything interesting about the concrete evolutionary history? That scientific theories and concepts should directly describe actual phenomena is a predominant view of science both in lay and scholarly circles. Under this conception logical positivists made it their primary task to define theoretical terms by the observable. In the same vein philosophers of biology have tried (not successfully in my view) to justify the concepts like homology or species by identifying necessary and sufficient conditions in terms of visible or directly verifiable features of organisms.

This apparently intuitive picture, however, has been criticized to be an overly simplistic view on the relationship between a scientific theory and reality (e.g. Suppes, 1967; Cartwright, 1983; Suppe, 1989). According to the critics the primary referents of scientific theories, concepts, and laws are not actual phenomena but idealized models. These models are not exact replicas of reality, but extract only certain features that are supposed to play essential roles in the scientific problem at hand. The present analysis is in line with this tradition. Causal graphs are highly idealized and thus possibly incomplete representations of complex causal interactions in living systems, but it is this idealization that affords explanatory power and general applicability. That is, on the condition that a model extracts the common causal structure of a population can it be used to predict the population's evolutionary trajectory or consequences of hypothetical interventions.

Most of these models, however, are still idiosyncratic to particular populations — e.g. population geneticists usually build, customize, or parameterize their model for each study object.[3] Homology thinking aims at even higher generality: its core idea is that some distinct species or organs allow for the same treatment/model in the analyses of their evolutionary fate or physiological performance. A homology statement is a historical hypothesis as to why such a unified explanation is possible at all. That is, it justifies the use of the same causal model based on evolutionary history, i.e. by the descent of the

[3]Models of adaptive evolution, however, may be extrapolated to the same or similar environmental conditions. In this regard, the analogical thinking and homological thinking represent two distinct ways to generalize evolutionary models.

causal graph from common ancestry. Hence homology is far from "residual," but has a significant explanatory value in biology — it allows an extrapolation of an evolutionary or physiological model to other contexts, and thus provides a basis for the highest-level generality in biological sciences.

5 Conclusion

The concept of homology presupposes phenotypic units on which identity relationships can be defined. The present analysis identified these units with causal graphs representing developmental or behavioral mechanisms and defined homology as graph isomorphism over lineages. The advantage of this formal concept is that it acknowledges the distinctive role of the study of homology while suggesting its connection to the traditional population genetics framework. That is, it not only provides definite meanings to such concepts like constraints, evolvability, and novelty, but also presents homology as a historical account or justification of the generalizability of evolutionary or physiological models. This is paralleled with the shift in the ontological nature of what can be said to be homologous: homology is a relationship between theoretical models, rather than concrete biological entities such as organs. Hence the proper relationship between homology to actual biological phenomena is not instantiation, but representation. Once conceived in this way the metaphysical problem of the alleged essentialism fades away.

The new account of homology prompts empirical, theoretical, and philo-

sophical researches on various topics, including the study of novelty and evolvability, the interplay between Evo-Devo and population genetics, implications of developmental flexibility, and the generalizability of biological models, to name a few. Another interesting philosophical question not mentioned above is the possibility of extending the current approach to another vexing concept in evolutionary biology, namely *species*. If homology is a partial matching of the causal structures between distinct species, it is tempting to define species by the whole causal structure — so that two organisms belong to the same species if their entire ontogeny and life history are represented by the same causal graph. This is a big question that requires an independent analysis, but will be briefly discussed in the presentation if time permitted.

References

Amundson, R. (2005). *The Changing Role of the Embryo in Evolutionary Thought: Roots of Evo-Devo*. Cambridge University Press, New York, NY.

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, 61(1-2):127–148.

Brigandt, I. (2007). Typology now: homology and developmental constraints explain evolvability. *Biology & Philosophy*, 22(5):709–725.

Brigandt, I. (2009). Natural kinds in evolution and systematics: Metaphysical and epistemological considerations. *Acta Biotheoretica*, 57(1-2):77–97.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press, New York, NY.

Ghiselin, M. (1997). *Metaphysics and the Origin of Species*. State University of New York Press, New York.

Gilbert, S. F. and Bolker, J. A. (2001). Homologies of process and modular elements of embryonic construction. *Journal of Experimental Zoology*, 291(1):1–12.

Kluge, A. G. (2003). On the deduction of species relationships: A précis. *Cladistics*, 19(3):233–239.

Love, A. C. (2009). Typology reconfigured: From the metaphysics of essentialism to the epistemology of representation. *Acta Biotheoretica*, 57(1-2):51–75.

Müller, G. B. (2003). Homology: The Evolution of Morphological Organization. In Müller, G. B. and Newman, S. (eds.), *Origination of Organismal Form: Beyond the Gene in Developmental and Evolutionary Biology*, pp. 51–69. The MIT Press.

Otsuka, J. (2015). Using Causal Models to Integrate Proximate and Ultimate Causation. *Biology & Philosophy*, 30(1):19–37.

Otsuka, J. (2016). Causal Foundations of Evolutionary Genetics. *The British Journal for the Philosophy of Science*, 67(1): 247-269.

Pigliucci, M. and Müller, G. B. (2010). *Evolution: the extended synthesis*. MIT Press, Cambridge, MA.

Riedl, R. (1978). *Order in living organisms: a systems analysis of evolution*. Wiley, New York, NY.

Rieppel, O. (2005). Modules, kinds, and homology. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 304(1):18–27.

Suppe, F. (1989). *The semantic conception of theories and scientific realism*. University of Illinois Press.

Suppes, P. (1967). What is a scientific theory? In Morgenbesser, S. (ed.), *Philosophy of Science Today*, pp. 55–67. Basic Books, Inc., New York.

Suppes, P. (2002). *Representation and Invariance of Scientific Structures*. CSLI Publication, Stanford, CA.

True, J. R. and Haag, E. S. (2001). Developmental system drift and flexibility in evolutionary trajectories. *Evolution and Development*, 3(2):109–119.

Wagner, G. P. (1989). The biological homology concept. *Annu. Rev. Ecol. Evol. Syst*, 20:51–69.

Wagner, G. P. (2014). *Homology, Genes, and Evolutionary Innovation*. Princeton University Press, Princeton, NJ.

Wagner, G. P. and Misof, B. Y. (1993). How can a character be developmentally constrained despite variation in developmental pathways? *Journal of Evolutionary Biology*, 6(3):449–455.

Wagner, G. P. and Stadler, P. F. (2003). Quasi-independence, homology and the unity of type: a topological theory of characters. *Journal of theoretical biology*, 220(4):505–527.

Woodward, J. B. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York.

Serendipity: an argument for scientific freedom?

Baptiste Bedessem and Stéphanie Ruphy
Université Grenoble Alpes

PSA 2016, Atlanta

Abstract

The unpredictability of the development and results of a research program is often invoked in favor of a free, desinterested science that would be led mainly by scientific curiosity, in contrast with a use-inspired science led by definite practical expectations. This paper will challenge a crucial but underexamined assumption in this line of defense of scientific freedom, namely that a free science is the best system of science to generate unexpected results. We will propose conditions favoring the occurrence of unexpected facts in the course of a scientific investigation and then establish that use-inspired science actually scores better in this area.

1. Introduction

"I didn't start my research thinking that I will increase the storage capacity of hard drives. The final landscape is never visible from the starting point." This statement made by the physicist Albert Fert (2007), winner of the 2007 Noble Prize for his work on the giant magnetoresistance effect, expresses a very common belief, especially among scientists, about the unpredictable nature of the development and results of a research program. Such retrospective observations feed a type of "unpredictability argument" often invoked in favor of a pure, disinterested science led by scientific curiosity, in contrast with a use-inspired or applied science led by practical considerations. Polanyi gave a somewhat lyrical form of this kind of unpredictability argument in his classical essay "The Republic of Science" (1962). Science, says Polanyi (1962, 62), "can advance only by unpredictable steps, pursuing

problems of its own, and the practical benefits of these advances will be incidental and hence doubly unpredictable. … Any attempt at guiding research towards a purpose other than its own is an attempt to deflect it from the advancement of science… You can kill or mutilate the advance of science, but you cannot shape it." In Polanyi's view, claims about the unpredictable nature of scientific development go hand in hand with a plea for an *internal* definition of research priorities: a problem should be considered important in light of considerations internal to a field of scientific inquiry and not (at least not primarily) in light of external considerations, such as practical utility. The orientation of the inquiry by practical objectives is then deemed epistemically counter-productive and vain: one should not attempt to predict the unpredictable.

In response to this line of defense of free science, some authors emphasize the epistemic fecundity of use-inspired science (Stokes 1997, Wilholt 2006, Carrier 2004) showing that the presence of practical objectives does not run counter to the building of fundamental knowledge: more fundamental knowledge may be needed to achieve some particular practical ends. Industry research on the giant magnetoresistance effect in the 1990s is a telling example of research undertaken under considerable pressure to produce applicable results but which nevertheless produced, along the way, new fundamental knowledge (Wilholt 2006).

Our aim in this paper is to develop another line of defense of the epistemic fecundity of applied science, by challenging a crucial but often implicit assumption in the traditional defense of scientific freedom based on scientific unpredictability (such as Polanyi's or Fert's), namely the assumption that a free science is the best system of science to generate unexpected facts. But what are actually the conditions favoring the emergence of novelty in the course of a scientific investigation? This important issue has not received much epistemological

attention.[1] We will fill this gap by first distinguishing two kinds of unpredictability arguments often mixed when debating on scientific freedom, to wit, unpredictability as unforeseen practical applications and unpredictability as *serendipity* (cases, as we will explain in more details, where unexpected facts open up new lines of inquiry). Focusing on the latter, we will propose two conditions that favor the occurrence of unexpected facts in the course of a scientific investigation. In light of these two criteria we will then compare pure, disinterested science and applied science as regards their capacity to generate novelty.

2.Two types of unpredictability arguments

Appeals to the unpredictability of scientific results actually refer to various kinds of situations, which need to be clearly distinguished. First, the notion of unpredictability of scientific results can designate unforeseen practical applications of fundamental knowledge. Second, it can refer to a serendipitous dynamics of scientific progress: a line of research may sometimes lead to a totally unexpected, surprising result, which opens a new direction of inquiry. These two kinds of unpredictability give rise to *distinct* arguments in favor of scientific freedom, unfortunately often mixed in discussions about the relative merits of pure science and application oriented science.

2.1 Unpredictability as unforeseen practical applications

When unpredictability refers to unexpected applications, the argument is the following: freedom of research should be preserved since a free, disinterested science is needed to generate a reservoir of fundamental knowledge, which then can be used to develop

[1] Wilholt and Glimell (2011, 353) do touch upon this issue when discussing the link made by proponents of the autonomy of science between freedom of research and diversity of approaches favoring the epistemic productivity of science. But they just note that it is a strong assumption and do no further discuss its validity.

applications. This argument was typically developed by Vannevar Bush who appealed to the now classically called linear model of innovation:

> "Basic research leads to new knowledge. It provides scientific capital. It creates the fund from which the practical applications of knowledge must be drawn. New products and new processes do not appear full-grown" (1945, 20).

The development of the H-bomb in the frame of the Manhattan project is a paradigmatic case, also invoked by Bush: "basic discoveries of European scientists" (1945, 20) about the structure of the matter is what made possible the military application. Another frequently cited example of unpredictable application is the invention of the laser, a widely-used technological device nowadays, made possible by pure theoretical developments in quantum physics during the first half of the XXth century.

We will not in this paper discuss further this first version of the unpredictability argument. Let us just mention that its underlying linear model of innovation linking pure science and practical applications has already been challenged on several grounds by various authors (e.g. Brooks, 1994; Leydesdorff, 1997; Edgerton, 2004; Rosenberg, 1992). We rather want to focus on the second (and also widespread) type of unpredictability arguments, whose validity has been much less scrutinized.

2.2 Unpredictability as serendipity

This second type of argument appeals to unpredictability in the sense of *serendipity*: an unexpected observation or result opens up a new line of research leading to a fundamental discovery. A very well known historical episode illustrating such a serendipitous scientific dynamics is the invention of the first antibiotic by Flemings, after he had accidentally

observed the effect of a fungi (Penicilium) on bacteria colonies (Flemings, 1929). Also often cited is the discovery of radioactivity by Henri Becquerel (1896): when working with a crystal containing uranium, Becquerel noted that the crystal had fogged a photographic plate that he had inadvertently left next to the mineral. This observation led to the hypothesis that uranium emitted its own radiations. Another, perhaps less cited instance of serendipitous scientific dynamics is the discovery of the chemotherapeutic cisplatine molecule by scientists initially working on the effects of an electric field on bacteria growth (Rosenberg *et al*.,1967). They observed that cell division was inhibited because of the unexpected formation of a chemical compound with the Platinum atoms contained in the electrode. This chemical compound, which they named cisplatine, was then successfully tested as an anti-proliferative agent against tumoral cells.

When unpredictability refers to such serendipitous discoveries, freedom of research is defended on the grounds that scientists should be able to freely change the direction of their research or open up new lines of inquiry, in order to be able to follow up on unexpected results, thereby generating new knowledge (which in turn will possibly lead to new applications). But to properly work as an argument favoring free, disinterested research over applied research, this "serendipity argument" actually presupposes that the occurrence of surprising facts is more likely to happen in the first system of science than in the second. For increasing the production of new knowledge (and possibly new applications) does not only depend on being able to freely follow up on unexpected facts, it also (obviously) depends on whether occurrences of unexpected facts are favored, to start with. Two types of considerations are thus mixed in the serendipity argument: considerations on the occurrence of unexpected facts and considerations on the (institutional, material) possibility to follow up on them.

We will not for the moment discuss the second type of considerations and focus on the first, which has been largely neglected in the literature on scientific freedom, namely the conditions that favor the occurrence of surprising facts. Our central issue is thus the following: is a use-inspired science less likely to generate unexpected results than a free science mainly fuelled by curiosity? After having clarified the notion of *unexpected* result, we will propose two criteria that, we will argue, favor the occurrence of such results and in light of which free science and applied science can be compared.

3. Conditions of emergence of unexpected facts

By "unexpected facts" occurring in the course of an inquiry, we simply mean here results (observations, outcomes of an experiment, etc.) that cannot be accounted for within the theoretical or, more largely, the epistemic framework in which the empirical inquiry has been conceived and conducted. This kind of "exteriority" is what leads scientists to move away from the initial explanatory framework and open up new lines of inquiry in search of an alternative one that could accommodate the unexpected results.

3.1 Isolation and purification of phenomena

It is now a well-known feature of contemporary experimental sciences that many of their objects under study are "created" in the laboratory rather than existing "as such" in the real world. When drawing our attention to this epistemologically important feature, Hacking (e.g. 1983, chap. 13) specified that we should not read this notion of "creation" of phenomena as if *we* were *making* the phenomenon, suggesting instead that a phenomenon is "created" in the laboratory to the extent that it does not exist outside of certain kinds of apparatus. This is typically the case for a phenomenon like the Hall effect: it did not exist "until, with great ingenuity, [Hall] had discovered how to *isolate*, *purify* it, create it in the laboratory" (Hacking

1983, 226, *our italics*). In other words, Hall created in 1879 the material arrangement - a current passing through a conductor, at right angles to a magnetic field –, for the effect to occur and "if anywhere in nature there [were] this arrangement, *with no intervening causes,* then the Hall effect [would] occur" (1983, 226, *our italics)*. Isolation, purification, control of intervening causes (i.e. control of physical parameters) are noticeable features of an experimental protocol that have a straightforward consequence directly relevant for our philosophical interrogation on serendipity: they tend to limit the number of causal pathways which can influence the response of the object or phenomenon under study experimentally. Unknown causal pathways existing in the real world are thus inoperant (or less operant) in laboratory conditions, thereby limiting the occurrence of unexpected results. Hence our first criterion to evaluate whether a certain system of science favors surprising results: the more the phenomena under study in that system are isolated, purified in highly regimented experimental conditions, the less likely the occurrence of unexpected results is.

Moreover, isolation, purification of phenomena often go hand in hand with another noticeable feature of laboratory sciences, described by Hacking as follows: "as a laboratory science matures, it develops a body of types of theory and types of apparatus and types of analysis that are mutually adjusted to each other" (1992, 30). In particular, a given theoretical framework determines the type of questions that can be probed experimentally, guides the design of apparatus and defines the type of data produced. Consequently, "data uninterpretable by theories are not generated" (Hacking 1992, 55). This process of mutual constraints is well illustrated for instance by recent experimental inquiries in particle physics, such as the quest for the Higgs Boson. Its existence was postulated in the frame of the Standard Model of theoretical physics (Higgs, 1964) and complex experimental apparatus have been developed with the explicit goal of "discovering" it (LEP, 2003). The "discovery" occurred in 2012 (ATLAS, 2012) but the high degree of tailoring of the apparatus to the

theory postulating the particle can be considered as imposing some kind of a priori structure on the phenomenon, so that particles such as the Higgs boson are not so much "discovered" than "manufactured" (Falkenburg, 2007, 53). In any case, the "discovery" of the Higgs boson was hardly a surprise and illustrates Hacking's more general contention about experimental inquiries typical of contemporary laboratory sciences as opposed to real-world experiments: "[their] results are more often *expected* than *surprising*" (1992, 37, *our italics*).

3.2 Theoretical unifying ambition

Another relevant characteristic of an experimental inquiry is the degree of generality of its theoretical framework. Scientists working within a theoretical framework with a large unifying scope will be reluctant to "leave" it and search for an alternative one when facing an unexpected result, and for good epistemological reasons: there is (obviously) a high epistemic cost of abandoning a theoretical framework that provides explanations for a large set of phenomena. The right move is rather to try to accommodate the surprising result by adopting, if necessary, *ad hoc* hypothesis or tinkering with some ingredients of the existing theoretical framework, so that the result looses its "exteriority" and ends up being integrated. And because of this well-known "plasticity" and integrative power of well-established theoretical frameworks with a large unifying scope[2], when a (at first sight) surprising result occurs, it rarely leads to the opening up of a new line of inquiry in search of an alternative explanatory framework, but rather gets integrated within the existing one, thereby losing its unexpectedness.

There is another reason why a high degree of theoretical generality does not favor the occurrence of unexpected results, which is linked to our previous remarks on the process of

[2] Classical references on these ideas of plasticity or integrative power are of course Kuhn's description (1962) of scientists being busy working on resolving anomalies in normal science and Lakatos' concept of "protective belt" of a research program (1978).

mutual adjustment between theoretical ingredients, apparatus and data. By constraining the type of experimental procedures developed and the type of data generated, a theoretical framework with a large unifying scope tends to *homogenize* the experimental works conducted to probe the various phenomena that it accounts for. And since a diversity of experimental approaches increases the possible sources of emergence of surprising facts, we can conclude that by reducing this diversity, theoretical generality makes the occurrence of unexpected facts less likely to happen.

The case of the etiology of cancer provides interesting illustrations of these two unexpectedness-diminishing effects of theoretical generality. The classical theory of cancer, the Somatic Mutations Theory (SMT), has been challenged for fifteen years or so by a new theoretical approach, the Tissue Organization Field Theory (TOFT) (Sonnenschein and Soto, 2000). First developed in the 1970's, the SMT rapidly became the dominant research theoretical framework on carcinogenesis (Mukherjee, 2010). This hegemony led to a high degree of homogenization of the experimental inquiries: the experimental procedures were all dedicated to the very standardized search for genetic mutations, in the context of molecular biology. Moreover, many, if not all surprising observations were made compatible with SMT by using *ad hoc* hypothesis (Soto, 2011). For instance, it was observed that various types of cancer were exhibiting large-scale disorganization of the genome. This observation was unexpected to the extent that it could not fit with SMT's fundamental postulate of punctual mutations. To integrate it in the frame of SMT, the existence of an original genetic instability of the cancer cells was then postulated (Rajagopalan, 2003).

4 Use-inspired science, pure science, and unexpected facts

In light of the criteria that we proposed above, how does pure, disinterested science score compared to applied science when it comes to favoring the occurrence of unexpected facts? A

helpful starting point is provided by Martin Carrier's insightful characterization of applied science:

> "Three methodological features can be observed whose combined or marked appearance tends to be characteristic of applied science: local models rather than unified theories, contextualized causal relations rather than causal mechanisms, real-experiments rather than laboratory experiment conducted for answering theoretical questions" (2004, 4).

4.1 Local models

Let us start with the contrast between local models and unified theories. Whereas pure science often aims at providing comprehensive and unifying theoretical frameworks (think of the Standard Model in particle physics or the Big Bang model in cosmology), use-inspired research is characterized by the coexistence of numerous local models, each determining the development of specific experimental procedures. An extreme case of this locality are for instance the design-rules used in the industry, which are built as laws guiding action (Wilholt, 2006). They are experimentally confirmed rules providing relations among different relevant parameters to manufacture industrial products. These rules are extremely specific: they apply to a very few number of situations and each of them determines a singular experimental practice. The use of local models is also widespread in the biomedical sciences, a typically use-inspired field of research. We will again draw on oncology to illustrate our point. Consider for instance the case of the development of radiotherapy protocols in the first half of the XXth century. The aim was to intervene on cancer to cure it, without any general model describing the mechanism of carcinogenesis. This program promoted the development of a variety of exploratory approaches using X-rays against cancer (Pinell, 1992). As there were

no standardized protocols, many experimental procedures were tested, changing the density of X-rays received, the distance of emission, the frequency of the radiotherapy sessions. In order to improve the efficiency of the therapeutic methods, scientists tried to build various local models describing the action of X-rays on cancer, corresponding to the variety of experimental procedures implemented. Grubbe (1949) formulated a model based on the inflammatory reaction to explain the effects of radiotherapy on cancer: the inflammation of the surrounding tissue beyond the effects of X-rays is responsible for the decrease of tumoral mass. This model is applicable to his specific use of X-rays: he applied very high doses, necessary to generate an inflammatory response. In parallel, Tribondeau and Bergonié, using more moderate doses, developed a model based on the proliferation of the cells in tumoral context, which led to the "Bergonié law": X-rays have a higher impact on proliferating cells (Tribondeau, 1959).

What lessons can be drawn from this first contrast between local models and unified theories? The answer is rather straightforward, given the link spelled out in the previous section between the level of generality of theoretical models and the occurrence of unexpected facts (our second criterion): by promoting the use of a diversity of local models and heterogeneous experimental protocols, applied science favors the occurrence of unexpected facts, whereas the penchant of pure science for comprehensive unifying theoretical frameworks, hence homogenized experimental protocols, does not.

4.2 Causal incompletness

Let us compare now pure science and applied science in light of our first criterion based on the degree of isolation and purification of the phenomena under study. A directly relevant feature of applied science is the use of what Carrier calls “contextualized causal relations” rather than full causal chains. Use-inspired science typically aims at directly intervening on a

process or phenomenon often disposing only of a partial knowledge of the causal chains involved and without being able to isolate it from various causal influences exerted by the rest of the physical world. A direct consequence of this feature of applied science is the low degree of control of its experimental protocols. By contrast, since pure science aims primarily at answering fundamental theoretical questions, it designs highly regimented experimental procedures that isolate and purify phenomena in order to be able to get empirical answers about the specific fundamental processes questioned in the theoretical investigation[3]. Moreover, building highly regimented experimental procedures requires knowledge of full causal chains in order to be able to better control the response of the system under study. The outcome of the application of our criterion is then again straightforward: compared with pure science, applied science favors the occurrence of unexpected facts to the extent that its experimental procedures are less controlled and based only on partial knowledge of the causal influences exerted on the phenomenon under study.

The etiology of cancer provides again interesting illustrations of our claim. Indeed, many current cancer therapies built in the frame of use-inspired research are based on contextualized causal relations. Typically, if a cellular agent is found to be massively expressed in cancer cells, drugs are designed to inhibit it, even if the whole causal chain determining its action is not known. For instance, a large amount of proteins promoting angiogenesis (the growth of blood vessels), notably VEGF (Vascular Endothelial Growth Factor), was found in tumoral cells, leading to the design of anti-VEGF molecules (Sitohy,

[3] Carrier sums up this contrast as follows: "Empirical tests often proceed better by focusing on the pure cases, the idealized ones, because such cases typically yield a more direct access to the processes considered fundamental by the theory at hand. But applied science is denied the privilege of epistemic research to select its problems according to their tractability (...). Practical challenges typically involve a more intricate intertwinement of factors and are thus harder to put under control". (2004a, 4) In the life sciences, this focus on "pure cases" means using "model organisms" or a limited number of well spread cell lines (e.g. the HeLa cells or the Saccharomyces Cerevisiae yeast) to elucidate fundamental biological mechanisms. And the use of such standardized objects tends to homogenize the experimental protocols.

2012). These molecules are used without considering the complete causal chain in which the VEGF is embedded. Only their known action on angiogenesis is considered. The clinical tests have led to unexpected observations: the use of an anti-VEGF molecule (Avastin) can stimulate tumor growth (Lieu *et al*., 2013)[4]. This example shows that the use of contextualized causal relations promotes the appearance of surprising facts by allowing unknown mechanisms to intervene in the experimental procedure.

5. Concluding discussion

Our previous analysis has established that several features of pure, disinterested science make it less hospitable than use-inspired science to the occurrence of unexpected facts. For all that, it does not follow that proponents of freedom of science cannot appeal anymore to unpredictability in the sense of serendipity to make their case. For the issue of which conditions favor the occurrence of unexpected facts is only half of the story. The other half is the possibility to actually follow up on these occurrences and open new lines of inquiry. And this other half raises different issues. What are the institutional, social structures of science that make it easier for scientists to re-orient their research when needed? To what extent an initial orientation of a scientific investigation by "external" practical needs is less compatible with the opening of new lines of inquiry than an initial orientation by epistemic considerations internal to the dynamics of a scientific field? When appealing to the serendipity argument,

[4] Interestingly, this observation led to new use-inspired research programs, aiming at identifing the molecular causal pathways giving rise to this tumoral resistance phenomenon. It has notably strongly oriented the research toward the precise understanding of the VEGF pathways (Moens, 2014). For instance, the study of the mechanisms of expression in cancer cells of various kinds of VEGF agents is becoming an important program of research (Li, 2014) and these works allow to build new fundamental knowledge about the action of the VEGF proteins.

proponents of free, disinterested science not only presuppose that it is the best system of science to generate unexpected facts to start with – a contention that we have challenged in this paper – but also that it actually gives more freedom to scientists to follow up on unexpected results. In other words, the issue of scientists' given possibility to change the direction of their research when needed is somewhat mixed, confused with the normative issue of what the aims of science should be (in short, increase knowledge following considerations internal to science *vs*. answer external practical needs). But it seems to us that the two issues should be kept separate. After all, one can very well conceive a system of science whose aims are primarily to answer society needs but which nevertheless leaves scientists free to choose the lines of inquiry that seem *to them* the most promising ways of fulfilling these needs (which includes changing research directions if needed). Otherwise put, one can very well conceive a use-inspired science which is not a *programmed* science in which scientists are asked to plan every step of their inquiry in order to achieve a given aim. And note that a pure, disinterested science may be as much programmed as a use-inspired science: the fact that scientists are left free to choose the aims of their research does not protect them from having to plan every step to reach these aims. In any case, our purport in this paper was not to attack pure, disinterested science. There are, no doubt, many good reasons to defend it, but the widespread, traditional one grounded on the unpredictability of scientific inquiry is certainly not the most epistemologically cogent and solid one.

REFERENCES

ATLAS Collaboration. 2012. "Obervation of a new particle in the the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". *Phy. Lett. B* 716(1) :1-29

Becquerel, Henri. 1896. “Sur les radiations émises par phosphorescence”. *Comptes-rendus de l'Académie des sciences*.

Brooks, Harvey. 1994. “The relationship between science and technology”. *Research Policy* 23(5):477-86

Bush, Vannevar. 1945. *Science, The Endless Frontier. A Report to the President by Vannevar Bush, Director of the Office of Scientific Research and Development*. Washington D. C.: National Science Foundation.

Carrier, Martin. 2004. “Knowledge and Control: On the Bearing of Epistemic Values in Applied Science”. In *Values and Objectivity in Science*, ed. P. Machamer and G. Wolters, 275-293. Pittsburgh, PA: University of Pittsburgh Press.

———. 2004a. “Knowledge gain and practical use: Models in pure and applied research”. In *Laws and Models in Science*, ed. D.Gillies, 1:17. London: King's College Publications

Edgerton, David. 2004. “The Linear Model Did not Exist. Reflections on the History and Historiography of Science and Research in Industry in Twentieth Century”. In *Science-Industry Nexus: History, Policy Implications,* ed. Karl Grandin and Nina Wormbs, 31-57. New-York: Watson.

Falkenburg, Brigitte. 2007. *Particles Metaphysics. A critical Account of Subatomic Reality*. Springer.

Fert, Albert. 2007. Interview published in *Le Monde,* October, 25, 2007.

Flemings, Alexander. 1929. “On the antibacterial action of cultures of a penicillium with special reference to their use in the isolation of b. influenza”. *J. Exp.Path*. 10:226-36.

Grubbe, Emil. 1949. *X-Ray Treatment: Its Origins, Birth, and Early History*. St.Paul and Minneapolis, MN: Bruce Publishing Company.

Hacking, Ian. 1983. *Representing and Intervening*. Cambridge University Press.

———.1992. “The Self-Vindication of the Laboratory Sciences”. In *Science as*

practice and culture, ed. A. Pickering, 29-64. The University of Chicago Press.

Higgs, Petter W. 1964. "Broken Symmetries and The Masses of the Gauges Bosons". *Phy.Rev. Lett* 13:508.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. The University of Chicago Press.

Lakatos, Imre. 1978. *The methodology of scientific research programs*. Cambridge: Cambridge University Press.

LEP Collaboration. 2003. "Search for the Standard Model Higgs Boson at LEP". *Physics Letters B* 565:61-75

Leydesdorff, Loet and Etzkowitz Henry. 1997. *Universities and the Global Knowledge Economy: A Triple Helix of University-Industry-Government Relation*. London, Cassel Academic.

Li, Dong et *al*. 2014. "Tumor resistance to anti-VEGF therapy through up-regulation of VEGF-C expression". *Cancer Lett* 346:45-52.

Lieu, Christopher H. *et al*. (2013). The association of alternate vegf ligands with resistance to anti-vegf therapy in metastatic colorectal cancer. *PLoS One* 8(10):e77117.

Moens, Stijn et *al*. 2014. "The multifaceted activity of VEGF in angiogenesis - Implications for therapy responses". *Cytokine Growth Factor Rev* 25:473-82.

Mukherjee, Siddhartha. 2010. *The Emperor of All Maladies. A Biography of Cancer*. * Scribner.

Pinell, Patrice. 1992. *Naissance d'un fleau. Histoire de la lutte contre le cancer en France (1890-1940)*. Métaillé.

Polanyi, Michael. 1962. "The Republic of Science: Its Political and Economic Theory". *Minerva* 1: 54-73.

Rajagopalan, Harith, Nowak Martin A, Vogelstein Bert and Langauer Christoph. 2003. "The

significance of unstable chromosomes in colorectal cancer". *Nat Rev Cancer* 3(9):695-701.

Rosenberg, Nathan. 1992. "Science and Technology in the Twentieth Century". In *Technology and Enterprise in Historical Perspective*. Oxford, Clarendon Press.

Rosenberg, Barnett *et al*. 1967. "The inhibition of growth or cell division in escherichia coli by different ionic species of platinum(iv) complexes". *J. Biol.Chem* 242(6):1347-5.

Sitohy, Basel. 2012. Anti-vegf/vegfr therapy for cancer: reassessing the strategies. *Cancer Res* 8:1909-14.

Sonnenschein, Carlos and Soto A- M. 2000. "Somatic mutation theory of carcinogenesis: why it should be dropped and replaced". *Mol. Carcinog*. 29(4):205-211.

Soto, Ana M. and Sonnenschein C. 2011. "The tissue organization field theory of cancer: a testable replacement for the somatic mutation theory". *Bioessays* 33(5):332-340.

Stokes, Donald E. 1997. *Pasteur's Quadrant. Basic Science and Technological Innovation*. The Brookings Institution.

Tribondeau, Jean B. 1959. "Interpretation of some results of radiotherapy and an attempt at determining a logical technique of treatment". *Radiation Research*, 11(4):587-588.

Wilholt, Torsten. 2006. "Design rules: Industrial research and epistemic merit". *Philosophy of Science* 73(1):66-89.

Wilholt, Torsten and Glimell H. 2011. "Conditions of Science: The Three-Way Tension of Freedom, Accountability and Utility". In *Science in the Context of Application*, eds. M.Carrier and A.Nordmann, 351-70. Boston Studies in the Philosophy of Science.

(Accepted for publication in *Philosophy of Science,*
subject to revision after presentation at 2016 PSA meeting)

Using Democratic Values in Science: an Objection and (Partial) Response[1]

S. Andrew Schroeder (aschroeder@cmc.edu),
Claremont McKenna College

draft of June 2016

Abstract

Many philosophers of science have argued that social and ethical values have a significant role to play in core parts of the scientific process. A question that naturally arises is: when such value choices need to be made, *which* or *whose* values should be used? A common answer to this question turns to political values — i.e. the values of the public or its representatives. In this paper, I argue that this imposes a morally significant burden on certain scientists, effectively requiring them to advocate for policy positions they strongly disagree with. I conclude by discussing under what conditions this burden might be justified.

1. Values in Science and the Political View

By now, most philosophers of science probably agree that there is an important place for so-called contextual (i.e. personal, ethical, political) values in core parts of the scientific process, especially in areas where science is connected to policy-making. Values may appropriately play a role in evaluating evidence (Douglas 2009), choosing scientific models (Elliott 2011), structuring quantitative measures (Reiss 2013, ch. 8; Stiglitz, Sen, and Fitoussi 2010; Hausman

[1] For comments on earlier drafts of this paper, I thank Alex Rajczi and the students in a seminar on science and values at Claremont McKenna College. For discussions on related topics, I thank Gil Hersch, Daniel Steel, and Branwen Williams. This work was supported in part by a research grant from the Claremont McKenna College Center for Innovation and Entrepreneurship.

2015), and/or in preparing information for presentation to non-experts (Elliott 2006; Hardwig 1994; Resnik 2001; Schroeder 2016). The natural follow-up question has received less sustained attention: when scientists should make use of values, *which* (or *whose*) values should they use?[2]

In some cases, philosophers of science criticize a value choice on substantive ethical grounds (e.g. Shrader-Frechette 2008; Hoffmann and Stempsey 2008). This suggests that the values to be used are the objectively correct ones. A second common view gives scientists latitude to choose whatever (reasonable) values they prefer or think best, usually supplemented by a requirement of transparency. This is suggested by many existing codes of scientific ethics, which impose few constraints on scientists in making such choices.[3] Finally, a third view says that scientists ought to use the appropriate political values — that is, the values held or endorsed by the public or its representatives — at least when those values are informed and substantively reasonable.[4] The most straightforward argument for this view grounds it in considerations of democracy or political legitimacy. If certain value choices are going to ultimately influence policy, then the public or its representatives have a right to make those choices (Douglas 2005; Intemann 2015; *cf*. Steele 2012; Kitcher 2001).

There are, of course further possibilities, and these views can be combined in more complex ways (e.g. requiring scientists to use political values in some domains, while permitting them to use their personal values in others). But if, for simplicity, we stick to these three primary

[2] In some cases, the justification for incorporating values into the scientific process dictates an answer. Feminist critiques of historically androcentric fields, for example, suggest that non-androcentric values are needed as a corrective. I set aside such cases in this paper.

[3] Mara Walli, Matthew Wong, and I discuss this at length in a work-in-progress.

[4] I set aside, then, cases where the values, say, of a policy-maker are unreasonable, in the sense that they lie outside the range of values that ought to be tolerated in a liberal society. In such cases, an advocate of the political view may permit or require scientists to reject those unreasonable values. (See e.g. Resnik 2001.) Also, in this paper I will set aside the important question of what the political view ought to say when the values of the public diverge from the values of policy-makers. The answer to this question, I think, will depend on one's theory of political representation.

options, I think the third, which I will call the *political view*, is the most attractive. More precisely, I think that in most cases where values are called for in core parts of the scientific process, scientists should privilege political values.[5] The most obvious concern with this view, and one that has received much attention from its advocates, is that it doesn't seem practical. It isn't feasible to ask citizens or policy-makers to weigh in at every point in the scientific process where values are required, and even if we could, non-experts often will not have the scientific background to fully understand the options before them. Substitutes for actual participation on the part of policy-makers or the public, such as asking scientists to predict what the public would choose or to determine what values policy-makers would hold upon reflection, seem to place unreasonable epistemic demands on scientists.

Douglas (2005), Intemann (2015), Guston (2004), and others have argued that these problems aren't insurmountable, by suggesting specific ways that the concerns of policy-makers and the public can be brought into the scientific process. And Kevin Elliott (2006; 2011) has suggested a more general way we might make progress. The political view goes hand-in-hand with a view of the relationship between science and policy that is widely-held: that the role of a scientist is to promote informed decision-making by policy-makers.[6] Bioethicists have extensively discussed how health care professionals can promote informed decision-making on the part of patients and research subjects. Theoretical and empirical research has led to a range of suggestions for how physicians can promote informed decision-making, even in cases where a patient's values may be uncertain, different research subjects may hold different values, and so

[5] This, of course, is proposed as a principle of professional ethics - not e.g. a legal requirement.

[6] See also Resnik (2001), Martin and Schinzinger (2010), and Schroeder (2016) for theoretical defenses of this idea, which is consonant with the mission statements of many scientific organizations and associations.

forth. Elliott's hope is that many of these suggestions can be adapted to the scientific case, or at least a parallel research program could be carried out, informed by the work of bioethicists.[7]

It is, of course, far from established that these proposals will work, but the range of options on the table strikes me as cause for optimism. And even if these solutions don't work in all cases, there is still bite to the political view, since it could still tell scientists to use political values *when they can determine those values*. Accordingly, in this paper I would like to describe a different and I think deeper concern with the political view, one which has been conspicuously absent from the literature thus far. In requiring scientists to guide certain aspects of their work by political values, we will sometimes in effect ask that they support political causes they may personally oppose and bar them from fully advocating for their preferred policy measures. We are, then, depriving scientists of important political rights possessed by the general public. In the remainder of this paper, I will spell out this objection more fully and explain why I think it has significant moral force. In the end, I will suggest that although there is reason to think that the objection doesn't ultimately undermine the political view, it nevertheless constitutes a significant cost that accompanies that view, which its proponents need to acknowledge.

2. Two Cases Where the Political View Seems Troublesome

The literature on values in science is vast and diverse, and so it will be useful to have some particular examples in mind. First, consider Douglas's (2000; 2009) argument that scientists should or must appeal to value judgments when resolving certain uncertainties that arise during the scientific process. Scientists conducting research into the potential carcinogen dioxin, for example, were faced with liver samples which had tumors that could not clearly be

[7] See also Schroeder (2016) for how this might go.

categorized as malignant or benign. In resolving such borderline or ambiguous cases, Douglas argues that scientists should appeal to contextual values, when the constitutive norms of science don't dictate any resolution. In this case, health-protective values would lead scientists to classify borderline samples as malignant; while concerns about overregulation would lead scientists to classify those same samples as benign (Douglas 2000).

Second, consider the many choices that scientists have to make when preparing their results for presentation. How should uncertainty be characterized? (Should 90% or 95% confidence intervals be used?) Which study results should be highlighted? (Which drug side effects should be discussed at length, and which included as part of a long list?) How should statistics be summarized? (As means or medians? Should results be broken down by gender, or presented only in aggregate?) In making choices like these, scientists frequently must appeal to values — to decide, for example, which pieces of information are important and which are not.[8]

It is, I presume, fairly uncontroversial that these value choices — how to resolve uncertainties in the research process and how to present results — can influence policy in foreseeable ways. Douglas, for example, argues that this is the case in the dioxin studies. Classifying borderline samples as malignant will make dioxin appear to be a more potent carcinogen, likely leading policy-makers to regulate it more stringently (2000, 571). Keohane, Lane, and Oppenheimer (2014) show how a presentation choice made by the Intergovernmental Panel on Climate Change led to poor policy outcomes, which likely could have been avoided by presenting information differently. More generally, we know from a wealth of studies in psychology and behavioral economics that the way information is presented to someone can strongly influence her subsequent choices (Thaler and Sunstein 2008), and there have been

[8] For discussions, see Elliott (2006), Hardwig (1994), Keohane, Lane, and Oppenheimer (2014), Resnik (2001), and Schroeder (2016).

several influential commentaries calling for scientists to more carefully "frame" their results (Nisbet and Mooney 2007; Lakoff 2010). So it seems straightforward that the value choices made by scientists can predictably affect policy.

If these value choices can influence policy, then in directing scientists to make them in accordance with political values — as opposed to the scientists' personal values — we are asking scientists to characterize policy-relevant material in a way that may promote an outcome they strongly disfavor. For example, suppose the scientists in Douglas's dioxin study value public health much more than they value keeping industry free from overregulation, but the public and its elected representatives have the opposite view. Further, suppose both views are substantively reasonable, in that they are within the range of policies eligible for adoption through democratic processes. In this case, the political view would tell the scientists to categorize borderline samples as benign, since that would better cohere with the public's values. This could make dioxin appear to have minimal carcinogenic effects, predictably leading to less regulation than would have occurred had the scientists classified borderline samples according to their own, health-protective values. Similarly, suppose an environmental economist conducting an impact study of a proposed construction project is herself deeply committed to the preservation of natural spaces. Nevertheless, if the public is strongly committed to economic development, the political view would require her to put front-and-center a detailed breakdown of the economic consequences of construction, while describing the ecological costs more briefly or in a less prominent place — likely frustrating her desire for preservation.

Notice that the concern here is not simply that scientists are being asked to provide information that will lead to an outcome they disfavor. I take it that any reasonable approach to scientific ethics will require that scientists communicate honestly, even in cases where that

promises to yield policies they don't like. Similarly, I presume that scientists must also be forbidden from presenting information in ways that, though technically accurate, are nevertheless misleading. The problem here is that Douglas's scientists are being asked to characterize results in one way (as benign) that could, *with equal scientific validity*, have been characterized differently (as malignant). And our environmental economist is being asked to present her results in one way (highlighting economic benefits), when an alternate presentation (one highlighting ecological costs) would be equally honest, accurate, objective, transparent, clear, and so forth. In each case, then, we have a collection of underlying data which can be described or characterized in different ways, neither of which appears to be more scientifically valid than the other. The political view insists that scientists choose the description grounded in values they don't accept and which seems likely to promote policy outcomes they disfavor. In this respect, the political view requires scientists to in effect advocate for, or at least tilt the playing field towards, political views they disagree with.[9]

3. Elliott and The Principle of Helpfulness

This seems clearly to be a significant imposition on scientists and thus a cost of the political view. It is therefore surprising that, so far as I can tell, philosophers who have argued for the political view have not commented on it. This is most striking in Elliott's work. Elliott, recall, argues that scientists should aim to promote informed decision-making among policy-makers, in something like the way physicians should aim to promote informed decision-making among patients. Standard accounts in bioethics say that it is the patient's values that carry the

[9] Can't we let the scientists advocate for their preferred positions in other ways? We could let scientists present their preferred interpretation separately. But if the political view is to have bite, presumably these alternate results will have to be clearly designated so and offered in a less prominent place (e.g. in an appendix or online supplement). And we should of course permit scientists to advocate for their views outside of their scientific papers/reports. But it seems likely that these (private) statements will carry much less policy weight than their scientific ones.

day: in normal cases, the physician's job is to help a patient make decisions that cohere with her own values. If the scientific cases is analogous, then the scientist's job is to help policy-makers make decisions that cohere with their (or the public's) values. This, in turn, suggests that scientists should use political values when resolving uncertainties, presenting results, and so forth. In other words, Elliott's proposal seems to imply the political view.[10]

The main defense Elliott offers for this view, however, relies on Scanlon's "Principle of Helpfulness":

> Suppose I learn, in the course of conversation with a person, that I have a piece of information that would be of great help to her because it would save her a great deal of time and effort in pursuing her life's project. It would surely be wrong of me to fail (simply out of indifference) to give her this information when there is no compelling reason not to do so.[11]

Elliott sums up the idea this way: "[I]n situations where one can significantly help another individual by engaging in an action that requires little sacrifice, it is morally unacceptable not to help" (2011, 139). If the political view, however, requires characterizing data or presenting information in ways that promote policy choices a scientist strongly opposes, then this Principle doesn't apply. When the pro-health scientist is required to classify ambiguous samples as benign, that does involve a sacrifice. A refusal to do so — which would hinder the pro-industry policy-maker's ability to make an informed regulatory decision — would not be done "simply out of indifference". It would be done out of the scientist's desire to protect public health.

---

[10] In some work, Elliott appears to suggest that transparency about values may be enough (Elliott and Resnik 2014). That is, he doesn't seem to place (many) constraints on scientists' value choices, so long as they are open about those choices. If that is Elliott's view — and it is not clear to me that it is — it strikes me as in tension with his insistence that scientists promote informed decision-making. Surely I can better help you make a decision that coheres with your values by working from your values, rather than by working from my own values (even if I am open about what I am doing). Further, even if scientists are open about their value choices, policy-makers frequently won't have the technical expertise to be able to reinterpret a scientific study, replacing one set of values (the scientist's) with another (their own). (If values could so easily be swapped out by non-specialists, then much of the debate about values would be unimportant. Transparency is all we would require.)

[11] Scanlon (1996, 224), quoted in Elliott (2011, 139).

(Similar things, obviously, can be said about the environmental economist asked to highlight the economic aspects of a proposed construction project.)

Scanlon's Principle of Helpfulness is a quite weak one, applying only in cases where the agent in question can put forward no significant burden of compliance. That Elliott uses it to justify his informed decision-making framework, and implicitly the political view, suggests that he thinks that such a view doesn't impose significant burdens on scientists. But if what I've said has been correct, that is wrong. Even if the political view is justified — and, as I've said, I think it is — we need to recognize that it asks a lot of scientists in cases where their values diverge from those of the relevant political body.

4. Physicians vs. Scientists

This, however, brings up an interesting question. If Elliott is right that the scientific case is analogous to the biomedical case, then shouldn't informed consent requirements in medicine be treated as similarly burdensome? Few bioethicists, though, would have sympathy for a physician who claimed that seeking informed consent constituted a significant ethical burden. (They may have sympathy for the claim that seeking informed consent is burdensome in more mundane ways — e.g. too time-consuming — but those complaints seem very unlike the scientists'.) I think that there is an important difference between the cases, which will help us to more clearly understand why the scientist is often burdened in a way that carries moral weight, while the physician normally is not.

We can see this by constructing a case which seems to put a physician in a position like the scientist's. Consider Jane, a doctor who strongly believes that the end of life for terminal patients is greatly enhanced by effective pain management, even if doing so shorten's the

patient's life or impairs his consciousness. For this reason, Jane has chosen palliative care as her specialty, making it her life's work to help dying patients avoid unnecessary pain. One of her patients, John, has continually insisted that he wants to remain as lucid as possible, even if that means agony. As he lies here, in agony, Jane suspects that if she framed the information properly — highlighting a medication's ability to relieve pain, while downplaying its cognitive effects — she might be able to get John to accept it. And accepting the medication, Jane strongly believes, would be much better for John. Nevertheless, standard interpretations of informed consent forbid her from doing so. Knowing that John is especially concerned about lucidity, she is ethically bound to highlight that information when informing him of his options. Unsurprisingly, John declines the pain medication and experiences what Jane regards as an awful death — precisely the kind of thing she went into palliative care to prevent.

Like our pro-health scientist, Jane has been asked to present information in a way that ultimately frustrates her deeply-valued goals. But imagine Jane complains to the ethics board at her hospital, arguing that it is burdensome to ask her to highlight to John the effects of pain medication on lucidity, because doing so would frustrate her deeply-held values. This complaint doesn't strike me as at all compelling. Why? Because Jane's values shouldn't hold any sway over John's medical choices. John has the right to reject pain medication, whatever Jane (or just about anyone else) thinks about it. Put another way, John has no obligation to take Jane's wishes into consideration, when he makes his decision. His decision is ultimately *his*.

Now, imagine our pro-health scientist complains to her ethics committee, asserting that it is burdensome to ask her to present her data in a pro-industry light, when it could with equal scientific validity be presented in a pro-health light, because doing so would frustrate her deeply-held concern for public health. Or imagine the environmental economist complaining about

having to foreground the economic benefits of the proposed construction project, since doing so will make it more likely that the project is approved and another natural space will be bulldozed. If we assume that the scientists are citizens of the society in question, then their situation is different from Jane's. As citizens in a democracy, their views should hold some sway over their government's policy choices. A government does have an obligation to take its citizens' views into consideration when making policy decisions. And when the government ultimately acts, it does so on the scientists' behalf. The decision is, in part, the scientists'.

The scientists, then, are stakeholders and even part-decision-makers in the associated policy-decisions, in a way that Jane is not a stakeholder in John's decision. This is true even if Jane cares more about John's decision than our scientists care about the policy decisions. We can see, then, that the political view isn't burdensome simply because it directs scientists to promote or advocate for outcomes they disfavor. It is burdensome because it sometimes directs scientists to promote or advocate for disfavored views, on matters that they have a right to speak on, to a body that purports to act on their behalf. This is what gives their burden its moral significance.[12]

5. Justifying the Burdens of the Political View

Some scientists have recognized the burdens that even neutrality — let alone the political view — would impose on them.

> Conservation biology is inescapably normative. Advocacy for the preservation of biodiversity is part of the scientific practice of conservation biology. If the editorial policy of or the publications in [the journal] *Conservation Biology* direct the discipline toward an "objective, value-free" approach, then they do not educate and transform society… To pretend that the acquisition of "positive knowledge" alone with avert mass extinctions is misguided… Without openly acknowledging such a perspective,

[12] What about cases where the scientists are not citizens of the society in question? In some cases, we can still make out a stakeholder claim. (When it comes to climate change, for example, we are all stakeholders in U.S. climate policy.) But such cases raise complications which I unfortunately can't discuss in a short paper like this one.

> conservation could become merely a subdiscipline of biology, intellectually and functionally sterile and incapable of averting an anthropogenic mass extinction. (Barry and Oelschlaeger 1996)[13]

Most conservation biologists enter that field because of a strong commitment to the value of biodiversity and the preservation of nature (Marris 2006). Similar things are surely true of other scientific disciplines. (My experience has been that public health researchers and economists studying inequality disproportionately share certain political values.) To the extent that these values diverge from the values of the public and its representatives, the political view would require these scientists to continually characterize their results in ways structured by a value system they find unacceptable. (In this respect, things would be quite different for, say, climate scientists. Although their work is controversial, it nevertheless is founded on values that are widely shared. The potentially catastrophic consequences of climate change are ones that virtually everyone cares about. Climate change deniers typically object to the *empirical* claims made by climate scientists - not to the basic values they hold.)

Is it fair, then, to tell a conservation biologist, who perhaps entered the field because of her love for natural spaces and has spent the bulk of her life collecting information that she hopes can be used to preserve them, that she is nevertheless ethically bound to resolve uncertainties in her research in ways favorable to economic growth, or to present her results in ways that highlight the economic value (as opposed to, say, the private or aesthetic value) of undeveloped land? I don't have a full answer to this question — such an answer would require more empirical information, as well as a fuller discussion of political philosophy — but I think we can see how the argument would go. There are a range of situations in which we impose significant

[13] This article was followed by a collection of commentaries, most of which generally supported the authors' views. Similar proposals seem to crop up frequently among conservation biologists, and are generally endorsed by those in the field (Marris 2006).

restrictions on speech and advocacy for people in important social positions. The Code of Conduct for U.S. judges, for example, bars judges from publicly endorsing candidates for political office and from making speeches for political organizations.[14] Uniformed U.S. military personnel are not permitted to participate in political fundraising, speak at political events, or display political signs, even on their private vehicles.[15] Other constraints on speech and advocacy seem ethically appropriate for politicians, police officers, lawyers, and others.

So, if there is an important public good served by constraining scientists' advocacy, it doesn't seem in principle problematic to do so. Two arguments along these lines seem promising. First, a distinctly political approach might argue that although imposing this burden on scientists does restrict important political rights of speech and advocacy, it is done in order to expand the political rights of others. By requiring scientists to work from the values of the public, the ability of the public to make informed policy choices and to effectively advocate for their own positions is enhanced. Thus, although the political view constitutes a loss of political freedom to scientists, that loss is more than balanced by the gain in political freedom to the public as a whole. (A view like this seems generally consistent with an approach to democracy like Brettschneider's (2007).)

Second, a straightforwardly consequentialist argument could point out the terrible consequences that threaten to follow if the public and/or policy-makers distrust scientific results. One of the primary arguments that has been put forward in favor of informed-consent approaches in bioethics has been that it promotes trust on the part of patients. Similarly, Elliott's informed decision-making approach — which implies the political view — seems like a promising way to

[14] http://www.uscourts.gov/judges-judgeships/code-conduct-united-states-judges

[15] http://www.dtic.mil/whs/directives/corres/pdf/134410p.pdf

promote trust in science (Elliott 2011, 133-6; *cf*. Hardwig 1994; Resnik 2001). If, then, the political view proves to be an effective way of promoting public trust in science, which in turn heads off the problems that ensue when policy-makers disregard science, that could justify imposing significant burdens on scientists.

Neither of these defenses, of course, is anywhere near complete. But both do strike me as quite reasonable, and so I don't think the concerns I've discussed in this paper should lead proponents of the political view to give up that position. That said, it is important to note the form that these defenses take. Neither attempts to show that the burden on scientists is not morally significant (as, perhaps, we might be inclined to say about the complaint of the palliative care physician). Instead, they each point to compensating benefits — not necessarily enjoyed by the scientists in question — which morally outweigh the scientists' burden. This means that the political view, even if it is justified, comes at a real cost to scientists, which is something its proponents need to acknowledge.

References

Barry, Dwight and Max Oelschlaeger. 1996. "A Science for Survival: Values and Conservation Biology," *Conservation Biology* 10: 905-11.

Brettschneider, Cory. 2007. *Democratic Rights: The Substance of Self-Government*. Princeton University Press.

Douglas Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67: 559-79.

Douglas, Heather. 2005. "Inserting the Public Into Science." In *Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making*, ed. Sabine Maasen and Peter Weingart, 153-169. Springer.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

Elliott, Kevin C. 2006. "An ethics of expertise based on informed consent." *Science and Engineering Ethics* 12: 637-61.

Elliott, Kevin C. 2011. *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Rsearch*. Oxford University Press.

Elliott, Kevin C. and David B. Resnik. 2014. "Science, Policy, and the Transparency of Values." *Environmental Health Perspectives* 122: 647-50.

Guston, David. 2004. "Forget Politicizing Science. Let's Democratize Science!" *Issues in Science and Technology* fall 2004.

Hardwig, John. 1994. "Toward and Ethics of Expertise." In *Professional Ethics and Social Responsibility,* ed. Wueste, 83-101. Roman and Littlefield.

Hausman, Daniel. 2015. *Valuing Health: Well-Being, Freedom, and Suffering*. Oxford University Press.

Hoffman, George and William Stempsey. 2008. "The Hormesis Concept and Risk Assessment: Are There Unique Ethical and Policy Considerations?" *BELLE Newsletter* 14: 11-17.

Intemann, K. 2015. "Distinguishing between legitimate and illegitimate values in climate modeling." *European Journal for Philosophy of Science* 5: 217-32.

Keohane, Robert O., Melissa Lane, and Michael Oppenheimer. 2014. "The ethics of scientific communication under uncertainty." *Politics, Philosophy & Economics* 13: 343-368.

Kitcher, Phillip. 2001. *Science, Truth, and Democracy*. Oxford University Press.

Lakoff, George. 2010. "Why it Matters How We Frame the Environment." *Environmental Communication* 4: 70-81.

Marris, Emma. 2006. "Should conservation biologists push policies?" *Nature* 442: 13.

Martin, Mike and Roland Schinzinger. 2010. *Introduction to Engineering Ethics (2nd ed.)*. McGraw-Hill.

Nisbet, Matthew and Chris Mooney. 2007. "Framing Science." *Science* 316: 56.

Reiss, Julian. 2013. *Philosophy of Economics: A Contemporary Introduction*. Routledge.

Resnik, David. 2001. "Ethical Dilemmas in Communicating Medical Information to the Public." *Health Policy* 55: 129-49.

Scanlon, Thomas. M. 1996. *What We Owe to Each Other*. Harvard University Press.

Schroeder, S. Andrew. 2016. "Communicating Scientific Results to Policy-Makers." Paper presented at the American Philosophical Association Conference (Pacific Division). Available at <http://apa-pacific.org/framed/download.php?file=200.pdf>.

Shrader-Frechette, Kristin. 1994. *Ethics of Scientific Research*. Rowman and Littlefield.

Shrader-Frechette, Kristin. 2008. "Ideological Toxicology: Invalid Logic, Science, Ethics About Low-Dose Pollution." *BELLE Newsletter* 14: 39-47.

Steele, Katie. 2012. "The Scientist qua Policy Advisor Makes Value Judgments." *Philosophy of Science* 79: 893-904.

Stiglitz, Joseph E., Amartya Sen, and Jean-Paul Fitoussi. 2010. *Mis-measuring Our Lives: Why GDP Doesn't Add Up*. The New Press.

Thaler, Richard and Cass Sunstein. 2008. *Nudge*. Yale University Press.

Two Roads Diverge in a Wood: Indifference to the Difference Between 'Diversity' and 'Heterogeneity' Should Be Resisted on Epistemic and Moral Grounds

Anat Kolumbus*, Ayelet Shavit* and Aaron M. Ellison

,,,
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference

from *The Road Not Taken*, by Robert Frost (1916)

Abstract:

We argue that a conceptual tension exists between "diversity" and "heterogeneity" and that glossing over their differences has practical, moral, and epistemic costs. We examine how these terms are used in ecology and the social sciences; articulate a deeper linguistic intuition; and test it with the *Corpus of Contemporary American English (COCA)*. The results reveal that 'diversity' and 'heterogeneity' have conflicting rather than interchangeable meanings: heterogeneity implies a *collective* entity that *interactively integrates* different entities, whereas diversity implies *divergence*, not integration. Consequently, striving for diversity alone may increase social injustice and reduce epistemic outcomes of academic institutions and governance structures.

* Equal main contributors.

Key words: collectivity, diversity, ecology, heterogeneity, injustice, institutional diversity.

Acknowledgments: We deeply thank the many different scholars, from very different disciplines, whose work and ideas helped us develop the ideas in this paper. In particular we want to mention Tal Israeli, Tamar Sovran, Nadav Sabar, Daryl G. Smith and Elihu Gerson. They all responded to a single email from an anonymous B.A. student with the same rigor, enthusiasm and respect as to an established full professor, and thus demonstrated the true spirit of academic inclusiveness this paper seeks to explicate. AS's work is supported by Tel Hai College and the ISF (Israeli Science Foundation) grant 960/12 and AME's work on diversity, heterogeneity, and inclusivity in science is supported by the Harvard Forest, and by grant DBI 14-59519 from the US National Science Foundation..

1. Introduction: Diversity in the Ecological and Social Sciences

The concepts of diversity and heterogeneity are two basic types of dissimilarity that are implicitly and commonly assumed to hold interchangeable meanings by scholars and laymen alike. However, when we examined their actual usage, a surprising conceptual discrepancy – in fact a tension – emerged. In this article we call attention to this tension between 'diversity' and 'heterogeneity'[1] and we argue that there are non-trivial epistemic, moral, and practical costs to science and society when this difference is glossed over.

Our critical examination is part of a large body of literature on the benefits of diversity for science and society. There exist strong epistemic (Shrader-Frechette 2002; Longino 2002; Solomon 2006b) and moral (Haraway 1979; Fricker 2007; Douglas 2009, 2015) arguments for diversity in institutions, governance structures, and ecological systems

1 In this article, we use the analytic tradition of concept notation. If quoting the concept's usage, it will appear as "X" (e.g., Fisher's "diversity" is defined as…), when explicitly mentioned as a concept it will appear as X (e.g., the concept of diversity is…), and when implicitly mentioned as a concept it will appear as 'X' (e.g., 'heterogeneity' here describes…).

("ecosystems"). For example, empirical evidence shows that diversity improves academic performance (Gurin et al. 2004; Freeman and Huang 2015; Page 2014), because diverse individuals hold different values (Longino 1990; Harding 1991), situated knowledge (Haraway 1989), socio-gender locations (Code 2006), research styles and specialities (Gerson 2013) and conflicting theoretical scaffolds (Wimsatt and Griesemer 2007). There also are costs associated with diversity, including feelings of isolation and alienation leading to reduced academic achievements of minorities (Armor 1972; Holoien 2013) and unbridgeable disagreements among researchers that disintegrate research groups (Gerson 2013; Shavit and Silver, accepted for publication).

There also are societal costs of divergence between scientists and non-scientists. Within the social realm, increased divergence from scientific worldviews may facilitate public manipulation by spreading ignorance – agnotology (Proctor and Schiebinger 2008) – and untrue and/or unjust environmental outcomes (Shrader-Frechette 2002). Within the scientific realm, divergence exempts scientists from responsibility for not assessing carefully enough social risks of generalizing their recommendations outside the laboratory, field, or model (Douglas 2009). Given the increasing science-society divergence, it is often non-experts who engage with the public – e.g., journalists teaching politicians about climate change or students teaching the underprivileged – which further widen the separation and may also silence local knowledge (Fricker 2007), e.g. by leading experienced mothers not to consider their comprehensive understanding and information as 'knowledge' compared to a young psychology student who never held a child, or depriving those living all their life near a spring to "know" their local flow rate compared to an

ecology student or governmental regulator who read published results taken at random from nearby streams (Shavit, Kolumbus and Silver, accepted for publication).

Given the fine line between the costs and benefits of constructive and destructive dissimilarities, interrogating the most basic concepts and measurements of dissimilarity seems important and timely. This paper aims for a step in that direction.

2. Definitions of Dissimilarity

Fundamental to both diversity and heterogeneity is the concept of "variance" (Fisher 1918, 1925). Briefly, measurable properties ("variables") of a group of individual entities (a "population" of cells, organisms etc.) are rarely identical. Rather, they will take on a range of values $\boldsymbol{y} = \{y_1, y_2, y_3, \ldots y_n\}$, where the value of the variable measured for the i^{th} individual is denoted y_i. When graphed as a histogram (Tukey 1977), these values are distributed, with the most frequent values clustered around the most common one and rarer values towards the edges.

The average value of the distribution of the measured variables (its expected value $E(\boldsymbol{y})$ or its mean value $\acute{y}$), equals the sum of all the individual measurements divided by the number of individuals, n: $\acute{y} = \sum_{i=1}^{i=n} \frac{y_i}{n}$. The variance, or "spread" of the distribution is the sum of the squared differences between each individual measurement and the mean: $\sigma^2 = \sum_{i=1}^{i=n} (y_i - \acute{y})^2$. The standard error of the mean $\left(\frac{\sqrt{\sigma}^2}{n}\right)$ provides intuitive estimates of how variable the set of measurements is. Under reasonable assumptions, ≈63% of the

measurements fall within ± 1 standard error of the mean, and ≈95% fall within ± 2 standard errors of the mean.[2]

In statistics (and hence in nearly all the social and natural sciences), means and variances are characteristics of single populations (groups of measurements), but heterogeneity usually is a composite property of a group of measurements taken from more than one population. For example, the classic analysis of variance (ANOVA) developed by Fisher (1918) is used to determine if two or more populations differ in their average measured traits (e.g., height). A basic assumption of ANOVA is that the variances of the populations being compared are equal; this is referred to as "homogeneity of variance" or "homoskedasticity". In contrast, if variances are unequal (heterogeneous or heteroskedastic), mathematical transformations of the data must be done to ensure that variances are homogeneous prior to comparing populations using ANOVA.[3] Note that 'heterogeneity' here describes only the variance as a problem to overcome in order to allow a *common basis* for comparison. Throughout the rest of this article, however, the concept of heterogeneity describes entities within a collective. "Diversity", if it is used at all in statistics, refers simply to describe a collection of datasets that describe a wide range of different, often incommensurate, variables.

In contrast, diversity is used widely in ecology (e.g., McGill et al. 2015) and the social sciences (e.g., Page 2011). Unlike variance or heterogeneity, diversity is not a simple, one-dimensional predicate. McGill et al. identified at least 15 different kinds of

2 Ellison and Dennis (2010) provide a full discussion of the assumptions behind these estimates and calculation of associated confidence intervals.

3 See Gotelli and Ellison (2012) for details and another example of a "cost" of heterogeneity.

ecological diversity; differences among them reflect the number of variables or populations that are measured (one or more), the spatial scale of measurement (local or regional), and whether it is measured within or between populations. Unlike 'variance' or 'heterogeneity' – both of which are interpretable on their own – 'diversity' has little meaning to an ecologist unless it is associated with an object. For example, the concept of *alpha* diversity refers to the number of different species in a locality, the concept of *gamma* diversity to the number of different species in a region [a collection of localities], and *beta* diversity measures population change between localities.[4]

In the social sciences, Page (2011) makes similar distinctions between three kinds of diversity: (1) *variation*, or diversity within a type, referring to quantitative differences in a specific variable; (2) *diversity of types*, referring to qualitative differences between types; and (3) *diversity of composition*, or the way types are arranged. Page's variation is directly analogous to an ecologist's alpha diversity, and his diversity of types and diversity of composition are analogous to different dimensions of an ecologist's beta diversity. Most social scientists use "diversity" as a catchall phrase not attached to any particular measured process (Page, personal communication), but we suggest that more attention should be paid to the dimensions of beta diversity.

Although 'diversity' appears to be used abstractly in common parlance and is implicitly assumed to mean something very similar to 'heterogeneity', when we examined deeply rooted linguistic intuitions of certain core examples, and tested these intuitions in large databases of linguistic usage, an interesting distinction between 'diversity' and

4 Each of these can be unweighted (i.e., simple counts of different species) or weighted by their abundance or sizes (Chao et al. 2014).

'heterogeneity' was revealed, with relevance for understanding and improving civil society and its institutions.

3. A Conceptual Tension Between Diversity and Heterogeneity

Whereas scientific language may seem indecisive or vague, artistic language can be precise and revealing. For example, Robert Frost's *The Road Not Taken* beautifully highlights diverging dimensions of a difference (i.e., 'diversity'), whereas the etymology of 'heterogeneous' implies something quite the opposite: an integration of multiple other (Gr.: *hetero*) kinds (Gr. *genus*) within a single whole.

We argue that attributing heterogeneity to something (e.g., a cell, computer, etc.) implies attributing an *integration* of mutual interactions among different entities that all belong to the same *collective*, whereas attributing diversity to a collection of objects or entities entails neither interactions nor a common collective.

An examination of English idiomatic constructions reveals clear distinctions in usage of diversity and heterogeneity. We would say that the parts of a cell or a clock are heterogeneous, but not that they are diverse. In contrast, we recognize a diverse collection of wall decorations or tools. There is an apparent semantic distinction here: cells and clocks are collectives whose functioning entails the integration of a number of interacting parts, whereas walls or garages function independently of the collection of items hanging on them. In other aspects of common usage, however, many objects in daily speech, including communities, populations, or universities, are called diverse or heterogeneous interchangeably.

The *Corpus of Contemporary American English* (henceforth: COCA; Davies 2008) provides a resource with which to examine common usage of diversity and heterogeneity in more detail. COCA contains more than 520 million words of texts, including scholarly

writing, fiction and nonfiction, newspapers and spoken recordings, and has tools to conduct complex searches for occurrences of words, phrases, parts of speech, other linguistic forms, and any combination thereof. Compilations of lists of co-occurrences (i.e., all types of words [adjectives, verbs, nouns, etc.] or specific words that appear near a target word) that can be used to infer intended meanings of predicates such as diverse or heterogeneous.

Sabar (2016) used COCA to infer motivations underlying regular co-occurrences of words. By identifying partial intersection of words that regularly co-occur more than expected by chance alone, Sabar identified *communicative strategies*: the choices of specific linguistic forms that best contribute to their intended message (e.g., "look" and "carefully" form the phrase "look carefully" that calls for visual attention). Thus, the generality of a communicative strategy that is evident in a particular example is established via a quantitative prediction of a non-random co-occurrence ("look" and "carefully" occur together and in sequence more frequently than expected by chance alone, and Sabar (2016) confirmed that "look" and "see" differ in meaning as a feature of attention by showing that "look" co-occurred more frequently with words such as "notice" than did "see").

We searched COCA and the *Wikipedia Corpus* (Davies 2015) for frequencies of "diverse" and "heterogeneous" and tested our hypotheses regarding differences in meaning between them using chi-square tests for non-random frequencies. "Diverse" occurred 12-30 times more frequently than "heterogeneous" in the corpora. In line with our hypothesis, "homogeneous", "collective", "whole", "integration" and "interaction" co-occurred significantly more frequently with "heterogeneous" than with "diverse" (improved prediction by, respectively, 58, 24, 8, 11, and 11%). Antonyms of these words ("single",

"individuals", "division", "separation") showed only random patterns of co-occurrence when they co-occurred at all (see tables 1-7 in the Appendix). A possible explanation for the latter findings is that while concepts of a collective whole seem to be more explicitly related to 'heterogeneity', words and meanings of singularity are relevant to both terms (in the case of heterogeneity they could relate both a single whole or to its parts). Nonetheless, it is evident that there is empirical support for our semantic intuition regarding 'heterogeneity' as interactions among diverse entities within a collective whole, and, perhaps more importantly, the empirical lack of a collectivist meaning for 'diversity'.

The attribute of diversity does not correctly describe collective entities because its meaning and reference are much wider than the concept of heterogeneity. A heterogeneous entity may be composed physically of nothing more than diverse entities, but as a collective, it entails multiple direct and indirect interactions, and feedbacks, among these entities. All reproducing biological groups (genomes, cells, metapopulations, etc.) are heterogeneous in the collective sense. Hence, additional information that refers to internal interactive processes improves models of heterogeneous entities and systems (Wade 1978; Roughgarden, accepted for publication). Some human groups – e.g., families, football teams or kibbutzim – would best be described as heterogeneous, whereas others – e.g., people waiting to pay the cashier – would not (Shavit 2008). There may be grave costs associated with failing to identify the goals of certain human groups as diverse or heterogeneous, as the next section portrays.

4. Illustrating the Diversity-Heterogeneity Trade-Off

4.1 Moral costs

Many – perhaps most – readers of this essay would say that promoting diversity is a social good because it is a stepping-stone to heterogeneity and thus to social justice. Although we may not yet have achieved a just and heterogeneous society, we should nonetheless promote diversity as much as possible and not dwell on the semantic particularities of distinguishing the concepts of diversity from heterogeneity. We think this line of thinking is misleading, and that the continuous focus on racial, ethnic, or gender 'alpha diversity' (i.e., headcounts) and use of the results of these measurements as a sufficient basis for discourse and policy, creates a vicious circle that may hinder social change in many of our institutions, in particular in our schools, colleges, and universities.

For example, in *Brown v. Board of Education* (1954), the Supreme Court of the United States ruled that segregation of African-American and Caucasian students in schools violated the Equal Protection Clause of the U.S. Constitution. One outcome of this decision was transporting students of different racial backgrounds into different school districts ("busing") to achieve diverse, "integrated" schools. This was intended to provide equal opportunities, academic aspirations, and achievements for all students and to improve relations among different races (Armor 1972). Unfortunately, according to some of its strongest supporters, busing did not improve academic aspirations or achievements (St. John 1975), sometimes decreased them and often worsened interracial relations: "integration … enhances ideologies that promote racial segregation, and reduces opportunities for actual contact between the races." (Armor 1972, 13).

In higher education, diversification is primarily done through "affirmative action". Many scholars support affirmative action (e.g., Bowen and Bok 2000; Rothstein and Yoon

2008), but others have argued that it leads to similar or worse outcomes than would have occurred in its absence (e.g., Sander 2004; Sander and Taylor Jr. 2012). For example, between 1988 and 2007, faculty of color made up only 17% of total full-time faculty, and that there had been little change in this number since the 1980's (Turner, González, and Wood 2008). Similar findings have been reported for the number of earned PhDs (NSF 2013).

However one thinks about affirmative action, we suggest that in the interest of promoting social justice that institutions should not measure diversity alone – how many people of different backgrounds are found at a certain time and place – nor wait for it "to work its magic" and reduce injustice. Smith (2015) identifies three problems with current mechanisms for promoting diversity in higher education: (1) responding to calls to improve diversity reactively rather than proactively, often by producing an internal quantified response to an external standardized requirement; (2) failure to include people from the many interacting parts of a university – faculty, staff, students, etc. – in discussions about diversity; and (3) making diversification into a specific program rather than an integral institutional function and goal. All of these common methods of "working towards diversity" are problematic precisely because they increase diversity but reduce heterogeneity. They track and magnify difference and divergence rather than encourage and enhance mutual interaction among all different co-occurring identity groups.

A more positive approach was reported by Walton and Cohen (2011), who conducted a very brief intervention in one's sense of social belonging (SOB) to a selective, largely Caucasian, college. After three years, there was a significant increase in the GPA (grade point average) of African-American students relative to control groups. SOB is central to a

heterogeneous community as it is a psychological aspect of being a part of an integrated collective.

We suggest that a trade-off exists between tracking diversity and building heterogeneity, which may result in a vicious circle leading to blaming those afflicted with social inequality for their under-representation. Since we are better at measuring discrete variables such as grades and gender than at measuring interactions such as SOB and research cooperation, we invest more effort in creating changes we can easily track rather than those that demand more complex, "beta type", measurements (e.g., institutional SOB, type of contacts with colleagues or task composition in the lab). As a result of neither measuring these latter dynamics nor investing in their visible change, alienation and lower academic achievements may persist among minority students and scholars (Syed, Azmitia, and Cooper 2011) even while their "diversity" increases. If this processes continues, a dangerous positive feedback may emerge, where not only will one's self-image and achievements be worsened, but also his/her social identity comes out worse than before affirmative action took place.

4.2. Epistemic Benefits

Aiming for heterogeneity rather than diversity often has epistemic benefits. Human collectives – as well as individual agents – have a variety of epistemic perspectives (Shavit, Kolumbus and Silver, accepted for publication). These perspectives differ in multiple inter-related ways, involve different backgrounds and experiences, and vary in ways of perceiving, explaining, and evaluating information about the world. Perspectives direct our attention to track a wide range of phenomena, promote diverse models to explain them (Griesemer 2014) and encourage adaptive-reflection by employing "…a variety of social perspectives, often…by taking the perspective of others" (Bohman 2006, 180).

Information is distributed asymmetrically between agents, so that some of it is known in general, some exclusive to certain groups, and some idiosyncratic to specific individuals (Sunstein 2003; Andesron 2006; Solomon 2006a; Gerson 2013); lack of interaction keeps pieces of information latent.[5] Diversity alone will not ensure that information is shared and provides fewer opportunities for agents to reflect on information that they can access only through interactions with others (Longino 2002; Tollefsen 2006).

Integrative working interaction across specialties – unlike the typical diverse-one-way adoption of ideas from one disciplinary to another – "includes coordinated efforts to pose and solve new research problems that can redefine specialty boundaries" (Gerson 2013, 516), and leads to developing new specialties. Tollefsen (2006) interweaves individual and collective knowledge in a way that demonstrates the benefits of epistemic heterogeneity. She suggested a framework of splitting a group that shares a common goal (e.g., works on a related set task or problems) into sub-groups; heterogeneity is manifested on an inter-sub-group level. Each sub-group is responsible for a different task, has its own sub-goals, and devises its own strategies and solutions. Mutual interactions result when the sub-groups return to the original group setting to present their suggestions and give feedback to other sub-groups. They encounter dissenting perspectives of out-groups and are forced to consider them and examine their own perspective closely. This self-scrutiny and actual encounters with critiques by other groups reveals problems, such as inaccuracies, leaps and gaps, and uncertainties, allowing the sub-groups and the integrated collective opportunities for self-correction (Tollefsen 2006).

5 There is an on-going discussion regarding the epistemic efficacy of deliberation, which is beyond the scope of this article.

Since all sub-groups are part of a larger community that shares a common goal, they both depend on other sub-groups and are depended upon by them. This framework is heterogeneous rather than diverse as the common goal and the inter-sub-group interactions serve to integrate the group. It also maintains differences, thus reducing the danger of group cohesiveness leading to unanimity and conformism, without promoting divergence. Such a framework increases the chances of achieving accurate results and obtaining a more just process of decision-making.

5. Conclusion

Diversity is not heterogeneity, and a continued focus on the former is not increasing the latter; instead, there is often a trade-off and tension between them. We illustrated how heterogeneity can better advance academic institutions and governess structures by integrating different people, identities, perspectives, and sources of information; it facilitates interactions among them, which have constructive epistemic and moral implications. Conversely, diversity alone often leads to divergence, is insufficient to resist social injustice and it misses epistemic opportunities that result from integrative working interactions. Institutions are often unaware of the diversity-heterogeneity tension or remain indifferent to it. They invest efforts in promoting diversity while neglecting heterogeneity, thus paying the costs of the trade-off and not reaping its benefits. Tracking alpha and disregarding beta diversity maintain this trade-off and obscures it. For moral and epistemic reasons we suggest noting this conceptual and practical difference and aiming for heterogeneity.

References

Anderson, Elizabeth. 2006. “The Epistemology of Democracy.” *Episteme* 3 (1-2): 8–22.
Armor, David J. 1972. “The Evidence on Busing.” *Public Interest* 28:90–126.
Bohman, James. 2006. “Deliberative Democracy and the Epistemic Benefits of Diversity.” *Episteme* 3 (3): 175–91.
Bowen, William G., and Derek Bok. 2000. *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions*. Princeton, NJ: Princeton University Press.
Chao, Anne, Nicholas J. Gotelli, T. C. Hsieh, Elizabeth L. Sander, K. H. Ma, Robert K. Colwell, and Aaron M. Ellison. 2014. “Rarefaction and Extrapolation with Hill Numbers: A Framework for Sampling and Estimation in Species Diversity Studies.” *Ecological Monographs* 84 (1): 45–67.
Code, Lorraine. 2006. *Ecological Thinking: The Politics of Epistemic Location*. *Ecological Thinking: The Politics of Epistemic Location*. Oxford, UK: Oxford University Press.
Davies, Mark. 2008. “The Corpus of Contemporary American English: 520 Million Words, 1990-Present.” Accessed February 15. http://corpus.byu.edu/coca/.
———. 2015. “The Wikipedia Corpus: 4.6 Million Articles, 1.9 Billion Words.” Adapted from Wikipedia. Accessed February 15. http://corpus.byu.edu/wiki/.
Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press.
———. 2015. “Politics and Science: Untangling Values, Ideologies, and Reasons.” *The ANNALS of the American Academy of Political and Social Science* 658 (1): 296–306.
Ellison, Aaron M., and Brian Dennis. 2010. “Paths to Statistical Fluency for Ecologist.” *Frontiers in Ecology and the Environment* 8 (7): 362–70.
Fisher, Robert A. 1918. “The Correlation between Relatives on the Supposition of Medelian Inheritance.” *Transactions of the Royal Society of Edinburgh* 52:399–433.

———. 1925. *Statistical Methods for Research Workers*. *Biological Monographs and Manuals*. Edinburgh: Oliver and Boyd.

Freeman, Richard B., and Wei Huang. 2015. "Collaborating with People like Me: Ethnic Co-Authorship within the US." *Journal of Labor Economics* 33 (3(S1)): S289–318.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford, NY: Oxford University Press.

Frost, Robert. 1916. *Mountain Interval*. New York, NY: Henry Holt.

Gerson, Elihu M. 2013. "Integration of Specialties: An Institutional and Organizational View." *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:515–24.

Gotelli, Nicholas J., and Aaron M. Ellison. 2012. *A Primer of Ecological Statistics. 2nd Edition*. Sunderland, MA: Sinauer Associates.

Griesemer, James R. 2007. "Tracking Organic Processes: Representations and Research Styles in Classical Embryology and Genetics." In *From Embryology to Evo-Devo*, ed. Manfred D. Laubichler and Jane Maienschein, 375–433. Cambridge, MA: MIT Press.

———. 2014. "Reproduction and the Scaffolded Development of Hybrids." In *Developing Scaffolds in Evolution, and Cognition*, ed. Linnda R. Caporael, James R. Griesemer, and William C. Wimsatt, 23–55. Cambridge, MA: MIT Press.

Griesemer, James R., and Michael J. Wade. 1988. "Laboratory Models, Causal Explanations and Group Selection." *Biology and Philosophy* 3 (1): 67–96.

Gurin, Patricia, Jeffrey S. Lehman, Earl Lewis, Eric L. with Dey, Sylvia Hurtado, and Gerald Gurin. 2004. *Defending Diversity: Affirmative Action at the University of Michigan*. Ann Arbor, MI: University of Michigan Press.

Haraway, Donna. 1979. “The Biological Enterprise: Sex, Mind, and Profit from Human Engineering to Sociobiology.” *Radical History Review* 20:206–37.

———. 1989. *Primate Visions: Gender, Race, and Nature in the World of Modern Science*. New York, NY: Routledge.

Harding, Sandra. 1991. *Whose Science? Whose Knowledge? Thinking from Women’s Lives*. Ithaca, NY: Cornell University Press.

Holoien, Deborah S. 2013. “Do Differences Make a Difference? The Effects of Diversity on Learning, Intergroup Outcomes, and Civic Engagement.” University Report, The University of Princeton.

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.

———. 2002. *The Fate of Knowledge*. Princeton, NJ: Princeton University Press.

McGill, Brian J., Maria Dornelas, Nicholas J. Gotelli, and Anne E. Magurran. 2015. “Fifteen Forms of Biodiversity Trend in the Anthropocene.” *Trends in Ecology and Evolution* 30 (2): 104–13.

National Science Foundation, National Center for Science and Engineering Statistics. 2013. “Survey of Earned Doctorates, 1998–2013 [NSF Publication No. 15-304].” Accessed February 19. http://www.nsf.gov/statistics/srvydoctorates/.

Page, Scott E. 2011. *Diversity and Complexity*. Princeton, NJ: Princeton University Press.

———. 2014. “Diversity without Silos: The Confluence of the Social and Scientific Teaching of Diversity.” *Independent School Magazine* 73 (4): 27–30.

Proctor, Robert N., and Londa Schiebinger, eds. 2008. *Agnotology: The Making and Unmaking of Ignorance*. Stanford, CA: Stanford University Press.

Rothstein, Jesse, and Albert H. Yoon. 2008. “Affirmative Action in Law School Admissions: What Do Racial Preferences Do?” Working Paper 14276, National Bureau of Economic Research.

Roughgarden, Joan. *Accepted for publication*. "Model of Holobiont Population Dynamics and Evolution: A Preliminary Sketch." In *Landscapes of Collectivity in the Life Sciences*, ed. Snait Gisis, Ehud Lamm, and Ayelet Shavit. Cambridge, MA: MIT Press.

Sabar, Nadav. 2016. "A Meaning Hypothesis to Explain Speakers' Choice of the Sign Look." PhD diss., City University of New York.

Sander, Richard H. 2004. "A Systematic Analysis of Affirmative Action in American Law Schools." *Stanford Law Review* 57 (367): 367–483.

Sander, Richard, and Stuart Taylor Jr. 2012. *Mismatch: How Affirmative Action Hurts Students It's Intended to Help, and Why Universities Won't Admit It*. New York, NY: Basic Books.

Shavit, Ayelet, Anat Kolumbus, and Yael Silver. *Accepted for publication*. "Epistemic Collectives, Heterogeneity and Injustice: The Case for Town Square Academia." In *Landscapes of Collectivity in the Life Sciences*, ed. Snait Gisis, Ehud Lamm, and Ayelet Shavit. Cambridge, MA: MIT Press.

Shavit, Ayelet, and Yael Silver. *Accepted for publication*. "To Infinity and Beyond!" Inner Tensions in Global Knowledge- Infrastructures Promote Local and pro-Active 'location' Information." *Science and Technology Studies*.

Shavit, Ayelet. 2008. *One for All? Facts and Values in the Debate over the Evolution of Altruism*. Jerusalem: Magness Press, in Hebrew.

Shrader-Frechette, Kristin. 2002. *Environmental Justice: Creating Equality, Reclaiming Democracy*. Oxford, UK: Oxford University Press.

Smith, Daryl G. 2015. *Diversity's Promise for Higher Education: Making It Work. 2nd Edition*. Baltimore, MD: Johns Hopkins University Press.

Solomon, Miriam. 2006a. “Groupthink versus The Wisdom of Crowds: The Social Epistemology of Deliberation and Dissent.” *The Southern Journal of Philosophy* 44 (1): 28–42.

———. 2006b. “Norms of Epistemic Diversity.” *Episteme* 3 (1): 23–36.

St. John, Nancy H. 1975. *School Desegregation: Outcomes for Children*. New York, NY: Wiley.

Sunstein, Cass. 2003. *Why Societies Need Dissent*. Cambridge, MA: Harvard University Press.

Syed, Moin, Margarita Azmitia, and Catherine R. Cooper. 2011. “Identity and Academic Success among Underrepresented Ethnic Minorities: An Interdisciplinary Review and Integration.” *Journal of Social Issues* 67 (3): 442–68.

Tollefsen, Deborah. 2006. “Group Deliberation, Social Cohesion, and Scientific Teamwork: Is There Room for Dissent?” *Episteme* 3 (1-2): 37–51.

Tukey, John W. 1977. *Exploratory Data Analysis*. New York, NY: Addison-Wesley.

Turner, Caroline Sotello Viernes, Juan Carlos González, and J. Luke Wood. 2008. “Faculty of Color in Academe: What 20 Years of Literature Tells Us.” *Journal of Diversity in Higher Education* 1 (3): 139–68.

Wade, Michael J. 1978. “A Critical Review of the Models of Group Selection.” *The Quarterly Review of Biology* 53 (2): 101–14.

Walton, Gregory M., and Geoffrey L. Cohen. 2011. “A Brief Social-Belonging Intervention Improves Academic and Health Outcomes of Minority Students.” *Science* 331 (6023): 1447–51.

Wimsatt, William C., and James R. Griesemer. 2007. “Reproduction Entrenchments to Scaffold Culture: The Central Role of Development in Cultural Evolution.”

In *Integrating Evolution and Development: From Theory to Practice*, ed. Roger Sansom and Robert N. Brandon, 227–324. Cambridge, MA: MIT Press.

Appendix

Table 1. Wikipedia Corpus total target words occurrences.

Diverse	**Heterogeneous**
30967	1096

Table 2. Co-occurrences of “heterogeneous”/ ”diverse” with “interaction”. Hypothesis: “heterogeneous”-“interaction” > “diverse”-“interaction”.

	Interaction present		*Interaction* absent	
	N	%	N	%
Heterogeneous	11	18	1085	7
Diverse	49	82	30918	93
Total	60	100	32003	100

P<.001

Table 3. COCA total target words occurrences.

Diverse	**Heterogeneous**
16685	1305

Table 4. Co-occurrences of “heterogeneous”/ ”diverse” with “collective”. Hypothesis: “heterogeneous”- “collective” > “diverse”- “collective”.

	Collective present		*Collective* absent	
	N	%	N	%
Heterogeneous	5	31	1300	7
Diverse	11	69	16674	93
Total	16	100	17974	100

P<.001

Table 5. Co-occurrences of “heterogeneous”/ ”diverse” with “whole”. Hypothesis: “heterogeneous”- “whole” > “diverse”- “whole”.

	Whole present		*Whole* absent	
	N	%	N	%
Heterogeneous	7	15	1298	7
Diverse	40	85	16645	93
Total	47	100	17943	100

P<.05

Table 6. Co-occurrences of “heterogeneous”/ ”diverse” with “integration”. Hypothesis: “heterogeneous”- “integration” > “diverse”- “integration”.

	Integration present		*Integration* absent	
	N	%	N	%
Heterogeneous	6	18	1299	7
Diverse	28	82	16657	93
Total	34	100	17956	100

P<.05

Table 7. Co-occurrences of “heterogeneous”/ ”diverse” with “single”. Hypothesis: “heterogeneous”- “single” < “diverse”- “single”.

	Single present		*Single* absent	
	N	%	N	%
Diverse	77	97	16608	93
Heterogeneous	2	3	1303	7
Total	79	100	17911	100

P>.05

Levels of Reasons and Causal Explanation

Abstract

My starting points are the claims that explanations are answers to why-questions, and that to answer the question why some event E occurred one must provide reasons why E occurred. The idea that all explanations of events are causal then becomes the theory that the reasons why some event occurred are its causes. My main thesis in this paper is that many "counterexamples" to this theory turn on confusing two levels of reasons. We should distinguish the reasons why an event occurred ("first-level reasons") from the reasons why those reasons *are* reasons ("second-level reasons"). An example that treats a second-level reason as a first-level reason will look like a counterexample if that second-level reason is not a cause. But second-level reasons need not be first-level reasons; nor (on my theory) need they be causes. Along the way I use the distinction between levels to diagnose the appeal of, and one main flaw in, the DN model of explanation.

1 A New Causal Theory of Explanation

It is obvious that some explanations of some phenomena speak of the causes of those phenomena. Simple examples come immediately to mind: the bridge collapsed because the wind reached a certain intensity, electrons flew off the metal because light shone on it. Much more controversial is the claim that *every* explanation of why some event happened must say something about the causes of that event. What's more, not only is it controversial whether this claim is true, it is also controversial how the claim should be understood. I have a new way of understanding the idea that all explanations of events invoke causes, one that, I think, is the most natural way to understand it. I also think that the idea, understood my way, is true (with one qualification[1]), and can be defended against the repeated claim that there exist non-causal explanations.

My theory starts with the idea, which has been held by many others, that explanations are answers to why-questions.[2] A theory of explanation, then, should say what it takes for a proposition to be an answer to a why-question. Now one standard form answers to why-questions take is "P because Q": "The tide is high because the moon is overhead" answers "Why is the tide high?" But there is another form answers to why-questions can take. The other form is "A/The reason why P is that Q."[3] Now because-answers and reasons-why answers are, in some sense, equivalent. "The tide is high because the moon is overhead" and "The reason why the tide is high is that moon is overhead" in some sense convey the same information. But I think that, for theoretical purposes, it is better to focus on reasons-answers. (I argue for this claim in (Skow 2016).)

A theory built around reasons-why answers will fill in the schema

[1] See footnote 6.

[2] Among those who hold that explanations are answers to why-questions are Hempel (1965)—with some qualifications, Bromberger (1992), and Van Fraassen (1980).

[3] I ignore here the forms used to give "teleological" explanations; I extend my theory to cover teleological explanations in (Skow 2016).

1. A reason why P is that Q iff . . .

What should the claim that "explanations of events are causal" look like, if put into the form (1)? Let "P" hold the place for a sentence that describes the occurrence of an event. (I won't try to say anything useful about which sentences do this.) Here is my proposal:[4]

(T) A reason why P is that Q if and only if the fact that Q is a cause of the fact that P.[5]

The same kinds of examples that lend credence to the idea that explanations of events are causal lend credence to its translation (T) into the language of reasons. The lighting of the fuse caused the bomb to go off; sure enough, it is also true that the reason why the bomb went off is that the fuse was lit. The electron's passing through a magnetic field caused it to accelerate; sure enough, the reason why it accelerated is that it passed through a magnetic field.

On the other hand, the same examples philosophers have thought are counterexamples to the idea that explanations of events are causal also threaten to be counterexamples to (T).

A bunch of these examples, I think, are based on the same mistake. There is a distinction to be made between "levels" of reasons. The examples fail because they confuse the two levels.[6] My aim in this paper is to introduction the distinction, and show how it can be used to defuse some examples. I will look, in particular, at Elliott Sober's claim that equilibrium explanations are non-causal, and Marc Lange's claim that "distinctively mathematical" explanations are non-causal (Sober 1981, Lange 2013).

[4]There are other theories of explanation that try to capture the idea that all explanations of events are causal—for example, (Salmon 1984) and (Lewis 1986). I do not have space here to explore the differences between their theories and mine.

[5]For stylistic convenience I sometimes speak of causation as a relation between facts, and sometimes as a relation between events. I remain neutral on which, if either, of these ways of speaking gets us closer to causation's "fundamental nature."

[6]I should say that there is one kind of counterexample that I think succeeds against (T): examples of "grounding" explanations. My true view is that every reason why a given event occurred is *either* a cause *or* a ground of its occurrence. But I will ignore grounding explanation in this paper.

2 Levels of Reasons

The distinction I want to introduce is that between

- a fact R being a reason why some event E occurred—then R is a "first-level" reason; and
- a fact F being a reason why R is a reason why E occurred—then F is a "second-level" reason, a reason why something else is a reason.

Reasons on the two different levels appear in answers to different why-questions. The first-level reasons are the facts that belong in the complete answer to the question *why E occurred*. The second-level reasons, on the other hand, belong in the answer to a different why-question: the question, concerning some reason R why E occurred, of *why R is a reason why E occurred*.

It is easy to come up with examples of first-level reasons. If I strike a match and, by striking it, cause it to light, then one reason why the match lit is that I struck it. What about an example of a second-level reason? We can find one by looking for the answer to the question of why the fact that I struck the match is a reason why the match lit. One answer (there are others) is: one reason why the fact that I struck the match is a reason why the match lit is that there was oxygen in the room at the time. In general, background conditions to a cause's causing its effect are, I hold, reasons why the cause is a reason why its effect happened. (Background conditions are not, however, the only kind of second-level reason; more on this in a bit.)

3 Second-Level Reasons Need Not Be First-Level Reasons

Here is the thesis about levels of reasons that I will defend in this paper:

> A fact can be a second-level reason without being a first-level reason. A fact F can be a reason why R is a reason why E happened, without F itself being a reason why E happened.

I say that F *need not* itself be a reason why E happened; I do not say that it *cannot*. The example I gave earlier shows that sometimes F *is* also a reason why E happened. The presence of oxygen,

besides being a reason why the striking of the match is a reason why the match lit, is also itself a reason why the match lit. But it is not always like this.

Here is an example in which a second-level reason is not also a first-level reason. Jill throws a rock at a window, Joan sticks out her mitt and catches the rock, and the window remains intact. The fact that Joan stuck out her mitt is a reason why the window remained intact. There is the first-level reason. *Why* is it a reason? The reason why it is a reason is that Jill threw a rock at the window. (You can test this with a counterfactual: if Jill hadn't thrown, certainly Joan's sticking out her mitt would not have been a reason why the window remained intact. The window wouldn't have "needed" Joan's help.) But this second-level reason is not also a first-level reason: that Jill threw a rock is *not* a reason why the window remained intact.[7]

In this case, the second-level reason that is not also a first-level reason is a fact that "corresponds" to the occurrence of an event: Jill's throwing of the rock. According to my theory (T), first-level reasons why events occur all correspond to events, since they are all causes. But not all second-level reasons are like the two examples we've seen so far (Jill's throw, the presence of oxygen); not all second-level reasons correspond to events.

In fact, I hold that laws of nature are second-level reasons that are not also first-level reasons. If I drop a rock from one meter above the ground, and it hits the ground at a speed of 4.4 m/s, the fact that I dropped it from one meter up is a reason why it hit the ground at 4.4 m/s. The law relating impact speed s to drop height d, namely $s = \sqrt{2dg}$ (assuming drag is negligible and d is small), is a second-level reason: it is a reason why my dropping the rock from one meter up is a reason why the rock was going 4.4 m/s when it landed. But it is not, in my view, also a first-level reason. It is not a reason why the rock is on the ground at 4.4 m/s.

Mentioning laws of nature probably brings to mind Carl Hempel's DN model of explanation, which says (I'm sure this is familiar) that an explanation of a fact F is a conjunction of facts that (i) entail F, and (ii) essentially contains a law among its conjuncts (Hempel 1965).

[7]This is also the kind of example many take to show that causation is not transitive; see for example (Hitchcock 2001).

Hempel's theory is not framed as a theory of the reasons why facts obtain, but it is natural to interpret it as committed to the thesis that whenever there are any reasons why some fact obtains, at least one of the reasons is a law of nature. I, along with many others, reject Hempel's theory, but I have a new diagnosis of where it goes wrong. Its mistake is to take certain second-level reasons, laws of nature, to also be first-level reasons.

I asserted without argument that laws are second-level reasons; but this is a natural view to have, on certain approaches to causation. One approach to causation takes laws to be central: whenever you have a cause and effect C and E, there are some laws connecting C to E—and C is a cause of E *because of* those connecting laws.[8] But that is just to say that whenever C is a cause of E, some law is a reason why C is a cause of E. Now I hold that when some fact F is a reason why C is a cause of E, then F is also a reason why C is a reason why E happened. So it follows from this theory of causation that laws are second-level reasons. If you start here, and in addition think that second-level reasons are always also first-level reasons, you head toward the characteristic thesis of the DN model, the thesis that among the reasons why some event happens is always at least one law. But this line of thought is fallacious, because second-level reasons need not be first-level reasons; and, on my view, laws that are second-level reasons are never first-level reasons.

I admit that I have given no direct argument that laws are not first-level reasons. I'd like to put the burden on the other side: why think the are? They are certainly second-level reasons: they are certainly reasons why causes are reasons why their effects happen. But as the Joan and Jill example shows, second-level reasons are not always first-level reasons. So why think they are in the case of laws? Certainly we have a sense that laws are "explaining something"; my view captures this sense, by assigning them the role of explaining why causes explain their effects. Why isn't that enough?

[8]Hempel endorses something like this idea about causation; see (Hempel 1965: 349). It has, of course, had many other defenders.

4 How The Levels Can Get Confused

I said that the flaw in the DN model is that it mis-classifies laws, which are second-level reasons, as first-level reasons. I also sketched an argument (with a false premise) that leads to this mis-classification: "laws are second-level reasons, and second-level reasons are always first-level reasons, so laws are also first-level reasons." But I'm not saying that Hempel or anyone else ever entertained this argument explicitly. Is there anything else to be said about how and why supporters of the DN model might have come to mis-classify laws as first-level reasons?

Yes, there is. Pragmatic effects, effects of the rules of conversation on information exchange, can produce "data" that misleadingly suggest that laws are first-level reasons.

The reasons why an event happened are the parts of the answer to the question of why it happened. So if we come across a conversation in which one person asks "Why did E happen?," and another person answers this question by citing some fact F; and if that answer strikes us as correct; then we have some good evidence that F really is a reason why E happened.

Some of the evidence that laws are (first-level) reasons why events happen appears to fit this pattern (but I will argue it does not). Imagine someone walks into the room just as the rock hits the ground at 4.4 m/s, and she sees that it hit at this speed (maybe the rock fell onto a device that measures impact speeds). A curious person, she asks me why it hit the ground at 4.4 m/s. I respond,

> Well, I dropped it from one meter up, and impact speed s is related to drop height d by the law $s = \sqrt{2dg}$ (and of course $\sqrt{2 \cdot 1 \cdot 9.8} \approx 4.4$).

Haven't I answered her question? And doesn't the law that $s = \sqrt{2dg}$ appear in my answer? If so, then the law is a reason why the rock hit the ground at 4.4 m/s—isn't it?

If the answers to these questions are "yes, yes, and yes," then, at least in some cases, a law is a reason why an event occurred. It's not hard to get from this conclusion to the claim (characteristic of the DN model) that this is so in *all* cases, and that when someone answers a

why-question *without* mentioning a law, her answer is incomplete.[9]

But the answers to these questions are not "yes, yes, and yes." To explain what I think is going on I need to introduce another distinction: the distinction between a *good response* to a question and an *answer* to a question. If someone asks a question, obviously one good way to respond is to answer the question. But not every good response is an answer.

A simple example suffices to establish this. Sally asks whether Caleb is coming to the party. I know he's supposed to go to the party. I respond by saying "He's sick." This is a good response. But it is not an answer. The only two possible answers are "yes (he's coming)" and "no (he's not coming)." I didn't say either of those things.

There is a theoretical reason why we should expect there to be good responses that are not answers. The notion of an answer is a semantic one. The relation between a proposition and a question, in virtue of which that proposition is an answer to that question, is a semantic relation. But the notion of a good response is a pragmatic one. Whether a response to a question is good is a matter of what a cooperative speaker should say. In some circumstances, a cooperative speaker should respond to a question by doing something other than, or something more than, answering the question. In the simple example, I know that if I just answer the question by saying "no," then Sally will immediately ask me why he's not coming. Since I can foresee that she'll ask that, and since I know the answer to this question too, I respond to her explicit question not by answering it, but by answering the expected follow-up question. It is okay in this case not to explicitly answer the question she asked, because what I do say, my answer to the expected follow-up, conversationally implies that the answer to her explicit question is no.

I did not, however, need to be so indirect. I could have responded by answering both questions. I could have said, "no, he's sick." Here my response is good, but again it contains information that is not part of the answer to the question she explicitly asked. What keeps it from being a bad response is that the additional information is relevant to the topic of our

[9]This "incompleteness" defense is most fully developed by Railton (1981). For one thorough argument against it, see (Woodward 2003: chapter 4).

conversation; and it is relevant because, though it is not an answer to her question, it is an answer to an expected follow-up question.

I think the same thing is going on in the dropped rock example. I responded to the question by saying

> Well, I dropped it from one meter up, and impact speed s is related to drop height d by the law $s = \sqrt{2dg}$.

My response is a good one, but (as we've seen) it does not follow that every part of my response is part of an answer to the question asked. In my view, the first part of my response—"I dropped it from one meter"—is an answer to the explicit question ("why did the rock hit the ground at 4.4 m/s?"), but the second part, the law, is not; it, instead, is an answer to an unasked follow-up why-question, a follow-up question I can anticipate would be asked immediately if I only answered the explicit question. The follow-up is, of course, why is the fact that I dropped it from one meter up a reason why it hit the ground at 4.4 m/s?

In summary: it is often a good thing to include a second-level reason in a response to the question why some event happened; but the fact that this is good thing to do is compatible with that second-level reason not being a reason why that event happened.

5 Equilibrium Explanations

I now have two distinctions: that between first- and second-level reasons, and that between a good response to a why-question an answer to a why-questions. The two together provide the key to defusing many problem cases for (T), the thesis that the reasons why something happened are its causes.

Elliott Sober argued that equilibrium explanations are not causal explanations. His main example of an equilibrium explanation was R. A. Fisher's answer to the question of why the ratio of males to females in the current adult human population is very close to 1:1 (Fisher 1931). "The main idea" of Fisher's answer, Sober reports, "is that if a population ever departs from

equal numbers of males and females, there will be a reproductive advantage favoring parental pairs that overproduce the minority sex. A 1:1 ratio will be the resulting equilibrium point" (201). Parents who overproduce the minority sex are likely to have more grandchildren. So if males outnumber females in the population, the fitter trait is to be disposed to have more female children than male; being the fitter trait, this disposition should increase in frequency, with the result that the sex ratio is pushed from male-biased toward equality. The opposite happens if females outnumber males. Now Sober claims that this is not a causal explanation, since

> a causal explanation...would presumably describe some earlier state of the population and the evolutionary forces that moved the population to its present configuration...Where causal explanation shows how the event to be explained was in fact produced, equilibrium explanation shows how the event would have occurred regardless of which of a variety of causal scenarios actually transpired. (202)

In other words: Fisher's explanation does not say, for example, that the sex ratio in the year 1000 was such-and-such, and that this caused the sex ratio in the year 1100 to be such-and-such, and so on. Instead it consists of a bunch of conditional facts: for each year in the sufficiently distant past, if the sex-ratio in that year had had any "non-extreme" value (non-extreme meaning not all males or females), then the sex ratio today still would have been 1:1.

The first thing I want to say is that Sober makes a claim about what the causes of the current sex ratio are that I reject. He thinks that the only relevant causes of the fact that the sex ratio is currently 1:1 are facts of the form *the sex ratio at time T is m:n*. I'm with those who reject this claim. The fact that the sex ratio in 1000 was m:n is "too specific" to be a cause of the current sex ratio. There is a less specific fact, the fact that the percentage of males in 1000 was not 0 or 100%, that is as well placed to be the cause. The less specific fact is "better proportioned" to the effect than the more specific one; so it gets to be the cause.[10]

[10]A "proportionality requirement" on causation is defended in Yablo (1992) and Strevens (2008). The claim that examples of explanations that, like Fisher's, abstract away from the nitty-

My disagreement with Sober might not seem to help much. Isn't Fisher's explanation still a counterexample to (T)? Even if the cause of the current sex ratio is that the sex ratio in the past was never extreme, Fisher's explanation doesn't cite this cause either; his explanation instead contains a bunch of other facts, namely the conditional facts described earlier. Doesn't it follow that these conditional facts, which are not causes, are reasons why the sex ratio is 1:1, and thus that (T) is false?

I deny that those conditional facts that Fisher offers up are reasons why the sex ratio is 1:1. But I can't just say this; for when Fisher offered those facts up in response to the question of why the sex ratio of 1:1, everyone celebrated his response, they did not reject it. How can his response be something to celebrate, if it didn't answer the question?

The distinctions I introduced earlier show why. Fisher's response was something to celebrate, because it was a *good response to the question.* But it can be a good response without containing an answer; in fact that's exactly what I think is going on.

I think that the reason why the sex ratio is now 1:1 is that the sex ratio in the past was never extreme. But this is not something anyone would believe, or even be able to come to know, without an accompanying answer to the question of *why* that is the reason. So a good response to the question of why the sex ratio is now 1:1 must include an answer to the question of why the fact that the sex ratio was never extreme in the past is a reason why it is 1:1 now. And *that's* the question that the conditionals in Fisher's response constitute an answer to. Those conditional facts are second-level reasons why some other fact is a reason why the sex ratio is 1:1.

---

gritty details of the causal process that produced the event being explained count as non-causal is repeated by Batterman in, for example, (Batterman 2000: 28) and (2010: 2). Batterman assumes that abstracting away from the details takes you away from the causes; but the proportionality requirement shows that in some cases at least this is not so. Less specific facts may be better proportioned to an effect than more specific ones.

6 "Distinctively Mathematical" Explanations

Marc Lange has recently described a class of explanations that he calls distinctively mathematical explanations, and argued that they are not causal explanations (Lange 2013). My interest is not in whether his examples qualify as non-causal by his criteria, but in whether they are counterexamples to (T). Here is one of the examples:[11]

> Why did a given person [say, Jones] on a given occasion not succeed in crossing all of the bridges of Königsberg exactly once (while remaining always on land or on a bridge rather than in a boat, for instance, and while crossing any bridge completely once having begun to cross it)?...[Because] in the bridge arrangement, considered as a network, it is not the case that either every vertex or every vertex but two is touched by an even number of edges. Any successful bridge-crosser would have to enter a given vertex exactly as many times as she leaves it unless that vertex is the start or the end of her trip. So among the vertices, either none (if the trip starts and ends at the same vertex) or two could touch an odd number of edges (488-89).

Here is what Lange says about why explanations like this one not causal explanations:

> these explanations explain not by describing the world's causal structure, but roughly by revealing that the explanandum is more necessary than ordinary causal laws are (491).

There is definitely something right, and deep, in what Lange says. But I do not think that his examples are counterexamples to (T).

Let P be the property of bridge-arrangements that a bridge-arrangement has if and only if either every land-mass or every land-mass but two is met by an event number of bridges. The (supposed) answer to the question of why Jones failed that Lange presents boils down to this:

[11]This example is also discussed in detail by (Pincock 2007).

(2) The bridges of Königsberg lacked P; and, necessarily, if a bridge arrangement lacks P, then no one can cross all the bridges exactly once.[12]

Now if (2) really is the answer to the question, then my theory is false. So is (2) the answer? There are two parts to (2). First is the fact that the bridges lacked P. Now it is no problem for my theory to recognize that this fact is a reason why Jones failed. For this fact is certainly a cause of his failure. The challenge to my theory comes if the second fact in (2) is a reason why Jones failed. For the second fact, that necessarily, no one can cross all the bridges exactly once, if the bridges lack P, cannot be a cause of Jones' failure.

I want to say the same thing about this example that I've said about the others. (2), I maintain, is not an answer to the question of why Jones failed. (2) contains an answer *as a part*—the fact that the bridges lacked P. But it has another part, the necessary truth, that is not part of the answer. How is this compatible with the evident fact that (2) is a really good thing to say in response to the question of why Jones failed? Because the part of (2) that is not an answer to this question *is* an answer to an obvious follow-up why-question, namely, why is it that the bridges' lacking P is the reason why Jones failed?

Lange's diagnosis of this example, and the others he discusses, is quite sophisticated, and I don't have the space here to go in to all the things he says about them. Let me at least, however, mention one further thing he says. At one point he writes, "Even if [these examples] happen to appeal to causes, they do not appeal to them as causes...any connection they may invoke between a cause and the explanandum holds not by virtue of an ordinary contingent law of nature, but typically by mathematical necessity" (496). I am quite taken by this idea that an answer to a why-question might appeal to causes but not appeal to them *as* causes. What might this mean, in terms of reasons why? Here is a natural suggestion: maybe in some cases a cause is a reason why its effect happened, but it is false that the *reason why* the cause is a reason why its effect happened is that it is a cause. The suggestion continues: cases like that are examples

[12]I'm going to take Lange's qualifications about always remaining on land etc. as given.

of "non-causal explanations."

I think the suggestion is plausible: if there truly are cases like that, they should be counterexamples to my theory. They are not, however, counterexamples to my theory as stated. I should amend my theory to make it more vulnerable:

(T2) A reason why P is that Q if and only if (i) the fact that Q is a cause of the fact that P, and (ii) the reason why the fact that Q is a reason why P is that the fact that Q is a cause of the fact that P.

Now the question is whether the Königsberg example, or any other example, is a counterexample to (T2). I have a lot of thoughts about this, but can only be brief here. Lange's idea is that since the "connection" between the bridges' lacking P, and Jones' failure, is secured by a mathematical truth (a theorem of graph theory), the bridges' lacking P, while a reason, is not a reason because it is a cause. I reject this claim. Even if the connection is secured by a mathematical truth, the cause is still a reason because it is a cause. This assertion requires defense, but I don't have the space to defend it here.

7 Conclusion

In this paper I have presented a new causal theory of explanation that says that the reasons why an event occurred are its causes. I also drew two distinctions: that between the reasons why E happened, and the reasons why those reasons are reasons; and that between an answer to a why-question, and a good response to a why-question. I used these distinctions to defend the theory against the claim that equilibrium explanations and distinctively mathematical explanations are non-causal; and I believe the distinctions can be used to defend it against a wide variety of other examples.

References

Batterman, Robert (2000). "Multiple Realizability and Universality." *British Journal for the Philosophy of Science* 51: 115-45.

— (2010). "On the Explanatory Role of Mathematics in Empirical Science." *British Journal for Philosophy of Science* 61: 1-25.

Bromberger, Sylvain (1992). *On What We Know We Don't Know.* The University of Chicago Press and CSLI.

Fisher, R. (1931). *The Genetical Theory of Natural Selection.* Dover.

Hempel, Carl (1965). "Aspects of Scientific Explanation." In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* The Free Press, 331-496.

Hitchcock, Christopher (2001). "The Intransitivity of Causation Revealed in Equations and Graphs." *The Journal of Philosophy* 98: 273âĂŞ299.

Lange, Marc (2013). "What Makes a Scientific Explanation Distinctively Mathematical?" *British Journal for the Philosophy of Science* 64: 485-511.

Lewis, David (1986). "Causal Explanation." In *Philosophical Papers, Volume II*. Oxford University Press.

Pincock, Christopher (2007). "A Role for Mathematics in the Physical Sciences." *Nous* 41: 253-75.

Railton, Peter (1981). "Probability, Explanation, and Information." *Synthese* 48: 233âĂŞ256.

Salmon, Wesley (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton University Press.

Skow, Bradford (2016). *Reasons Why.* Oxford University Press.

Sober, Elliott (1983). "Equilibrium Explanation." *Philosophical Studies* 43: 201-10.

Strevens, Michael (2008). *Depth.* Harvard University Press.

Van Fraassen, Bas C. (1980). *The Scientific Image.* Oxford University Press.

Woodward, James (2003). *Making Things Happen.* Oxford University Press.

Yablo, Stephen (1992). "Mental Causation." *The Philosophical Review* 101: 245-80.

In Defense of the Actual Metaphysics of Race

Abstract. In a recent paper, David Ludwig (2015, 244) argues that "the new metaphysics of race" is "based on a confusion of metaphysical and normative classificatory issues." Ludwig defends his thesis by arguing that the new metaphysics of race is non-substantive according to three notions of non-substantive metaphysics from contemporary metametaphysics. However, I show that Ludwig's argument is an irrelevant critique of actual metaphysics of race. One interesting result is that actual metaphysics of race is more akin to the metaphysics done in philosophy of science than mainstream analytic metaphysics.

1. Introduction

In David Ludwig's (2015, 44) recent article "Against the New Metaphysics of Race," he argues for the provocative thesis that "the new metaphysics of race" is "based on a confusion of metaphysical and normative classificatory issues." Furthermore, to continue to engage in such a "methodologically dubious metaphysics of race" is, in Ludwig's (2015, 262) opinion, "a bad idea." Key to Ludwig's critique is that he defines "metaphysicians of race" as "committed to the ideal of one fundamental ontology of race," much like other metaphysicians engaged in mainstream analytic metaphysics (Ludwig 2015, 245). Furthermore, for Ludwig, "the new metaphysics of race" consists of disputes about "one fundamental ontology of race" (Ludwig 2015, 245). In his critique, Ludwig focuses on two debates in the new metaphysics of race.

The first is the debate about whether races exist according to the one fundamental meaning of 'race' in current, ordinary English in the United States (Ludwig 2015, 257). I'll call this *the US race debate**.[1] According to Ludwig (2015, 251, 253, 256, 260), some interlocutors

[1] The asterisk is intentional. I'm calling this debate 'the US race debate*' because I think Ludwig has changed the focus of the relevant debate. I borrow the convention of using an asterisk to flag when the meaning of a term has been changed from Joshua Glasgow (2009, 140).

in the US race debate* are Anthony Appiah, Joshua Glasgow, Michael Hardimon, Sally Haslanger, Quayshawn Spencer, and Naomi Zack.

The second debate in the new metaphysics of race is about whether humans have races according to the one fundamental meaning of 'race' in the life sciences (Ludwig 2015, 254). I will call this *the biological race debate**. Ludwig (2015, 251, 253, 259) claims that, among others, the interlocutors of the biological race debate* are Robin Andreasen, Bernard Boxill, A.W.F. Edwards, Adam Hochman, Jonathan Kaplan, Koffi Maglo, Armand Leroi, Massimo Pigliucci, Neven Sesardic, and Alan Templeton.

Ludwig defends his thesis using an argument premised on the claim that the new metaphysics of race is non-substantive according to three notions of non-substantive metaphysics from contemporary metametaphysics: one from Eli Hirsch, one inspired from Theodore Sider, and one from Ludwig himself. The relevant background here is that recent metametaphysics has been preoccupied with what constitutes a "substantive" metaphysical dispute, which, roughly, is a dispute that is *really* about metaphysics as opposed to some other topic, like how we use language (Hirsch 2005, 67).

While I agree with Ludwig that to engage in a metaphysics of race that confuses metaphysical and normative classificatory issues is a bad idea, and while I think that the new metaphysics of race (as Ludwig defines it) might be based on such a confusion, I will show that the work that *actual* metaphysicians of race are doing involves no such confusion. In other words, the point of this paper is show that Ludwig's argument is an irrelevant critique of the actual metaphysics of race.

For clarity, by 'actual metaphysicians of race', I'm talking about the same group of scholars that Ludwig is talking about in his critique, and by 'actual metaphysics of race' I'm

talking about the same body of work that Ludwig is talking about in his critique.[2] However, unlike Ludwig (2015, 245), I will not require actual metaphysicians of race or actual metaphysics of race to be "committed to the ideal of one fundamental ontology of race," even with respect to a particular linguistic context.

I will begin by clarifying Ludwig's argument and his defense of each premise. Second, I will show that even if Ludwig's argument is a good critique of the new metaphysics of race, it's irrelevant to the actual metaphysics of race. Finally, I will provide closing remarks where, among other things, I will clarify how the actual metaphysics of race is more akin to the metaphysics done in the philosophy of science than mainstream analytic metaphysics. As for objections, I will respond to them along the way.

2. Ludwig's Argument and Its Defense

2.1 The Basic Argument

Though Ludwig does not state his argument explicitly, a charitable reconstruction of it is below:

(1) If the new metaphysics of race is non-substantive, then it is based on a confusion of metaphysical and normative classificatory issues.

(2) The new metaphysics of race is non-substantive.

(3) So, the new metaphysics of race is based on a confusion of metaphysical and normative classificatory issues.

[2] For instance, like Ludwig (2015, 244), I consider Joshua Glasgow to be an actual metaphysician of race, and, like Ludwig (2015, 263), I consider Glasgow's actual metaphysics of race to consist of work like his book *A Theory of Race* and his article "On the New Biology of Race."

Ludwig states (3) as his thesis in the first paragraph of his opening remarks.[3] Ludwig states (2) in his opening remarks as well and at several points throughout his paper.[4] Ludwig also treats (2) as a reason for adopting (3).[5] However, since there is a logical gap between (2) and (3), it's charitable to add (1) as a suppressed premise.[6]

2.2 Ludwig's Defense of His Premises

Though Ludwig takes the truth of (1) for granted, he offers three, in-depth defenses of (2) that utilize three different notions of non-substantive metaphysics. Ludwig's first defense of (2) is the following:

(4) The new metaphysics of race is substantive only if there is exactly one allowable and fundamental ontology of race for each of its race debates.

(5) If there is a plurality of legitimate biological subdivisions below the species level or a plurality of equally allowable specifications of 'race' for each race debate in the new metaphysics of race, then there is a plurality of allowable ontologies of race for each race debate in the new metaphysics of race.

(6) The antecedent of (5) is true.

(7) So, it's not the case that the new metaphysics of race is substantive.

Ludwig claims (4) in section 3.1 and justifies his constraint on substantive metaphysics from how he defines 'a metaphysics of *x*.' For Ludwig (2015, 245, 251), a project on the

[3] See Ludwig (2015, 244).

[4] See Ludwig (2015, 245, 260-262).

[5] See, especially, sections 3.1-3.3 and 4 in Ludwig (2015).

[6] [removed for blind review]

"metaphysics of *x*" assumes that metaphysicians of *x* are committed to "one fundamental ontology" of *x* that rules out "a plurality of equally allowable ontologies" of *x,* at least for the relevant linguistic context.[7] Since a substantive metaphysics of *x* must at least be a metaphysics of *x,* it follows that a substantive metaphysics of *x* requires exactly one allowable and fundamental ontology of *x.* Substituting 'race' for '*x*' gives us (4).

As for (5), Ludwig states that the first disjunct of (5)'s antecedent leads to (5)'s consequent in section 2. Here Ludwig (2015, 247) follows Kaplan and Winther (2013) in arguing that if there is a plurality of equally legitimate but distinct ways of subdividing species into "legitimate biological kinds," then "[e]mpirical evidence underdetermines the ontological status of race," which in turn, permits a plurality of allowable ontologies of race (Ludwig 2015, 246-247). In particular, Ludwig (2015, 245, 247-249) argues that "both racial realism and antirealism" are allowable ontologies of race given different equally legitimate ways of subdividing a species, and even in the same race debate. An example is how Zack (2002) uses the fact that humans have no subspecies to defend racial anti-realism in the US race debate*, while Spencer (2014) uses the fact that humans have a population subdivision that matches the current US census racial scheme to defend racial realism in the same race debate.

Ludwig states that the second disjunct of (5)'s antecedent leads to (5)'s consequent in section 3.1. In his words, "If there is a plurality of equally allowable specifications of 'race', there is also a plurality of equally allowable ontologies of race" (Ludwig 2015, 251). Interestingly, Ludwig never defends this assertion because he takes it to be obviously true.

[7] See Ludwig (2015, 251) for (4) and see Ludwig (2015, 245) for Ludwig's view on the metaphysics of *x.*

Next, Ludwig defends (6) by defending the truth of each disjunct in the antecedent of (5). As for the first disjunct, Ludwig (2015, 246-247) argues that there is a plurality of legitimate biological divisions below the species level (e.g. population subdivisions, monophyletic levels, subspecies, etc.) because, first, legitimate biological kinds are *interest dependent,* and, second, there is a plurality of "explanatory interests" among biologists in different research contexts (e.g. population genetics, phylogenetic systematics, etc.). As for the second disjunct, Ludwig reaches it by making an induction from what's going on in the two most popular race debates in the new metaphysics of race: which are the US race debate* and the biological race debate*.

Ludwig (2015, 254) argues that there is a plurality of equally allowable specifications of 'race' in the biological race debate* since biologists in different research programs use 'race' in different ways that suit their needs. For instance, Ludwig (2015, 254) points out that 'race' is often used as a synonym for 'subspecies' in systematic biology, but often used as a synonym for 'ecotype' in ecology. As for the US race debate*, Ludwig takes a more circuitous route to the conclusion that there is a plurality of equally allowable specifications of 'race' in that debate. First, Ludwig (2015, 255) appeals to Glasgow et al.'s (2009) empirical research on how Americans use 'race' to argue that 'race' is "polysemous" in the current US. Next, Ludwig (2015, 257-258) argues that the *context* for the US race debate* has not been "sufficiently specified" to narrow the debate to "exactly one fundamental candidate meaning of 'race' in the United States." Hence, according to Ludwig, from induction, the second disjunct of (6) holds as well.

Ludwig's second defense of (2) utilizes Hirsch's notion of non-substantive metaphysics. The second defense is below:

(8) A dispute is merely verbal if each side can plausibly interpret the other

side as speaking a language in which the latter's asserted sentences are true.

(9) A dispute is non-substantive if it is merely verbal.

(10) Each side can plausibly interpret the other side as speaking a language in which the latter's asserted sentences are true in the new metaphysics of race.

(11) Thus, the new metaphysics of race is non-substantive.

(8) is a direct quote from Ludwig (2015, 259), which is itself a summary of Hirsch's (2005; 2008) view on non-substantive metaphysics.

Hirsch defends his distinction between merely verbal disputes and ones that aren't with several examples from the history of science and philosophy. For instance, Hirsch (2005, 73) shows that the dispute among classical physicists about whether a projectile's final velocity is equal to its initial velocity on Earth was not a merely verbal dispute because physicists on both sides could not charitably interpret the other side's assertions as true. In other words, both sides were using the same meanings of 'projectile', 'velocity', 'Earth', etc., and what they disagreed about were the laws of motion. In contrast, Hirsch (2008, 407-408) shows that the dispute between John Locke and Joseph Butler about whether a tree can survive a change in its parts was merely verbal since either side could charitably interpret the other side's assertions as true using the other's meaning of 'identity'. In short, a merely verbal dispute for Hirsch is one where the disputants are either talking past one another or merely arguing about how we do (or should) use language.

As for (9), we can infer that it's a premise from how Ludwig (2015, 259-260) uses 'merely verbal' and 'nonsubstantive' at this point in his paper. Furthermore, Ludwig's

vocabulary here is uncontroversial since it's the same vocabulary that Hirsch (2005, 67) uses. As for (10), Ludwig endorses it when he says the following:

> Realists like Andreasen, Edwards, Leroi, Sesardic, and Spencer can interpret antirealists as speaking the truth in a language in which 'race' refers to subspecies, populations with visible traits that mark relevant biological differences, populations with cognitive differences, and so on. Antirealists like Glasgow, Lewontin, Hochman, Maglo, and Zack can interpret realists as speaking the truth in a language in which 'race' refers to genetic clusters, patterns of mating, clades, and so on (Ludwig 2015, 259-260).

Finally, Ludwig defends (2) in a third way using his interpretation of Sider's notion of non-substantive metaphysics. Ludwig's third defense of (2) is below:

(12) A dispute about an expression *E* is non-substantive if its disputants are endorsing multiple, equally joint-carving candidate meanings for *E.*

(13) The new metaphysics of race is a dispute that is non-substantive according to (12).

(14) The new metaphysics of race is non-substantive.

(12) is directly from Ludwig (2015, 261), and is a rough summary of Sider's (2011, 46-49) view of non-substantive metaphysics. Sider defends the non-joint-carving condition in his definition of 'non-substantivity' from his stipulation of what metaphysics is about.

For Sider (2011, vii) the "central task" of metaphysics is "to discern the ultimate or fundamental reality underlying the appearances." We are supposed to describe this reality using a privileged language, so-called Ontologese, which is privileged exactly because all of its expressions (e.g. terms, quantifiers, etc.) are "joint-carving," which means that they carve out the

world's fundamental structure (Sider 2011, vii).[8] So, naturally, when we find that one or more of the expressions that we've used to formulate a question *Q* does not have exactly one, best joint-carving meaning, it's likely that a debate about *Q* is not about the fundamental structure of the world, and thus, is not a substantive metaphysical debate in Sider's sense.

With that said, it's important to note that Ludwig's summary of Sider is rough, and does not reflect Sider's (2011, 49) "revised" definition of a non-substantive dispute. What Ludwig presents is Sider's unrefined view, which occurs at the beginning of section 4.2 in chapter 4 of Sider's *Writing the Book of the World*. However, later on in section 4.2, after Sider considers multiple problems with his unrefined view, he settles on what he calls his "revised" definition.[9] Nevertheless, since Ludwig uses Sider's unrefined notion of non-substantivity in his critique, that's what I'll focus on as well. However, for clarity, I'll say that (12) expresses *Sider-style non-substantivity* as opposed to Siderian non-substantivity.

In any case, Ludwig (2015, 261) asserts and defends (13) when he says that Spencer's, Leroi's, Pigliucci's, and Hochman's biological definitions of 'race' are all "equally joint-carving candidates" for 'race' because they are all "objective ways of distinguishing between populations below the species level." Furthermore, Ludwig (2015, 261-262) bolsters his support for (13) when he says that Hardimon's, Glasgow's, Feldman and Lewontin's, and Appiah's biological definitions of 'race' are also equally joint-carving candidates for 'race' because they are all "non-joint-carving" meanings.

3. Why Ludwig's Argument is an Irrelevant Critique of Actual Metaphysics of Race

[8] For Sider's clarification of "Ontologese," see Sider (2011, 171-173).

[9] For Sider's "revised" definition, see Sider (2011, 49).

Even though Ludwig has provided a valid argument that may be sound as well, it turns out that Ludwig's critique does nothing to undermine the actual metaphysics of race. The latter is partially because Ludwig's critique is not *about* the actual metaphysics of race, it's about a hypothetical metaphysics that he calls 'the new metaphysics of race'.

Remember that the new metaphysics of race is, according to Ludwig (2015, 245), and by definition, constituted by disputes about "one fundamental ontology of race." Furthermore, remember that Ludwig claims that people like Glasgow, Haslanger, Appiah, and Spencer are engaged in one such dispute, the US race debate*, and people like Andreasen, Pigliucci, Kaplan, and Templeton are engaged in another such dispute, the biological race debate*. However, these last two claims are simply false.

For one, the term 'fundamental ontology' is not even a phrase used in actual metaphysics of race. For instance, it does not appear *once* among the actual metaphysics of race that Ludwig (2015, 263-265) cites, and he cites 40 such publications. Second, some actual metaphysicians of race embrace a pluralist ontology for the nature of race in the relevant context. For example, at the beginning of Spencer's (2014, 1026) article on the "national" meaning of 'race' in the US, he concedes that ordinary Americans are using multiple "geographic" and "ethnic" meanings of 'race'. In fact, Spencer (2014, 1026) explicitly says, "Hence, I acknowledge upfront that there are several ways that Americans use 'race'."

However, Ludwig could object here. Specifically, Ludwig (2015, 257) interprets Spencer's focus on the national meaning of 'race' in the US as an endorsement of it being "the only relevant candidate meaning for philosophical debates about the referent of "race" in the United States." While the latter is a possible interpretation of Spencer's project, it's not the most charitable one given how he presents his project at the beginning of his article. Spencer (2014,

1025) begins by saying upfront that his project is merely "to debunk" the idea that "folk racial classification has no biological basis." Spencer attempts to accomplish that goal by showing that 'race', in its national meaning in the current US, is a directly referring term for a biological entity—a set of particular human populations—that presently happens to be biologically real in virtue of being a level of human population structure. Thus, given how Spencer (2014, 1026) presents his own project, his race theory is compatible with there being a pluralist nature of race in the current US context. Furthermore, this interpretation best explains why Spencer (2014, 1026) says that "there are several ways that Americans use 'race'."

There are other actual metaphysicians of race who embrace pluralism about the nature of race as well. For instance, Pigliucci and Kaplan (2003, 1162-1163) are happy to grant that both the ecotype and the subspecies are equally legitimate ways of dividing a species into biological races. It's just that they believe that humans have ecotypes, but not subspecies. In fact, Pigliucci and Kaplan (2003, 1163) explicitly say, "Races, then, can be defined and picked out in a number of ways."

Finally, there are plenty of actual metaphysicians of race who do not embrace pluralism about the nature of race, but who do entertain pluralism as a metaphysical possibility, which is enough to show that they do not presuppose that there is a single fundamental ontology of race in the relevant context. For instance, after obtaining messy results about how ordinary Americans use 'race' and race terms in a widely distributed survey, Glasgow (2009, 75) entertains the possibility that ordinary Americans are sometimes "talking past each other" when they use 'race', much like we sometimes do when we use 'jade'. In fact, Glasgow (2009, 75) explicitly says, "So maybe 'race' is used in some contexts to refer to a social kind of thing and in other contexts to a biological kind of thing." That doesn't sound like somebody who presupposes that

there is a single fundamental ontology of race in the US context. Now, even though Ludwig's argument is not about actual metaphysics of race, it could still be a relevant critique of actual metaphysics of race. So to that I now turn.

In order to know whether Ludwig's argument succeeds in critiquing the actual metaphysics of race, we need to know more about the debates among actual metaphysicians of race. Clearly, the US race debate* and the biological race debate* are not debates among actual metaphysicians of race. However, the US race debate and the biological race debate are. *The US race debate* is the debate about the nature and reality of race according to what 'race' means in the ordinary discourse of contemporary Americans, but only when 'race' is used to classify humans. The latter debate actually exists because all of the individuals that Ludwig places in the US race debate* have expressed an interest in the focus I've just articulated.[10] *The biological race debate* is the debate about whether humans have any races in a nontrivial biological sense of 'race'. The latter debate actually exists as well.[11] These are the two race debates that Ludwig was attempting to critique, and given these distinctions, we can see that Ludwig's argument really isn't relevant to these two debates.

For one, neither the US race debate nor the biological race debate satisfies Hirsch's criterion for a non-substantive dispute. The US race debate is not a merely verbal dispute because racial realists in that debate, such as Haslanger and Spencer, cannot plausibly interpret racial anti-realists in that debate, such as Appiah and Glasgow, as speaking a language in which

[10] For evidence, see Appiah (1996, 42), Glasgow (2009, 15), Haslanger (2012, 133), and Spencer (2014, 1025).

[11] For evidence, see Andreasen (1998, 200-201, 205), Pigliucci and Kaplan (2003, 1161-1164), Maglo (2011, 362-363), and Templeton (2013, 262-263).

anti-realist race theories are true, and vice versa. For instance, if Glasgow (2009, 33) is correct about (H1*) being part of the non-negotiable semantic content of 'race' in the ordinary discourse of Americans, then Spencer (2014, 1026) is incorrect about 'race' directly referring to a set of human populations in the national racial discourse of Americans, and vice versa.[12] The biological race debate is not a merely verbal dispute either. For instance, if Pigliucci and Kaplan (2003, 1165) are correct that humans subdivide into "biologically significant" ecotypes, then Hochman (2013, 347) is incorrect that humans do not subdivide into "meaningful biological units," and vice versa.

Next, even if the US race debate or the biological race debate is non-substantive in a Ludwigian or Sider-style sense, that fact does not imply a "confusion about metaphysical and normative classificatory issues" as (1) claims. This is because actual metaphysicians of race are adopting a different view of *substantive* metaphysics—namely, one that does not require metaphysical disputes about race to presuppose a single fundamental ontology of race or anything about joint-carving. Thus, while Ludwig's argument is relevant to the hypothetical new metaphysics of race, it doesn't make contact with actual metaphysics of race.

Interestingly, when Ludwig defines 'the new metaphysics of race', he anticipates the worry that his focus on it may mischaracterize actual metaphysics of race. In response, Ludwig (2015, 245) says, "However, I do not want to engage in a verbal dispute about the meaning of "metaphysics of race"… this article only challenges a certain type of metaphysics of race while proposing an alternative deflationist and normative metaphysics of race." However, this reply is

[12] (H1*) is the claim that a race is, at least, a group of human beings that is distinguished from other groups of human beings by visible physical features, of the relevant kind, that the group has to some significantly disproportionate extent (Glasgow 2009, 33).

perplexing because if the new metaphysics of race is a purely hypothetical metaphysics that does not describe the disputes in actual metaphysics of race (as I've shown), and, in addition, if the disputes in actual metaphysics of race already do away with monist and fundamentalist assumptions about race (as I've shown), it's hard to imagine what the purpose is for lodging Ludwig's critique in the first place. In any case, we can rest assured that actual metaphysicians of race are immune to Ludwig's critique because they've already been vaccinated against monist and fundamentalist assumptions about race.

5. Closing Remarks

In this paper, I've shown that Ludwig's critique of the new metaphysics of race is irrelevant to the actual metaphysics of race. However, I've said little about the conditions of substantivity that actual metaphysicians of race adopt. In addition to the bare minimum of "not talking past one another" (Glasgow 2009, 28), actual metaphysicians of race embrace disputes about how certain linguistic communities actually use 'race' (e.g. Pigliucci and Kaplan 2003, 1162-1163; Glasgow 2009, 6), and embrace disputes about how certain linguistic communities should use 'race' (e.g. Haslanger 2012, 221-247; Hochman 2014, 80). However, actual metaphysicians of race do not embrace disputes that have unimportant social and scientific consequences. For instance, Haslanger (2012, 300) motivates the US race debate by pointing out that engaging in it will help us frame and evaluate social policies and appropriately address stubborn inequalities in health. Also, Pigliucci and Kaplan (2003, 1170) point out that engaging in the biological race debate can help biologists debunk hereditarian hypotheses about race and intelligence, yield insights into human evolutionary history, and yield insights into human migration history.

Interestingly, the criteria for substantive metaphysics that actual metaphysicians of race adopt make the metaphysical disputes in the actual metaphysics of race more akin to metaphysical disputes in the philosophy of science (e.g. the species debate, the nature of natural kinds, the ontic structural realism debate, etc.) than those in mainstream analytic metaphysics (e.g. debates about the nature of fundamentality, grounding, modality, substantivity, etc.). For instance, Matthew Slater's (2015) stable property cluster theory of natural kinds has a real shot at explaining why some kinds support epistemically reliable inductions in a domain while others don't, which could help systematic biologists achieve more agreement about how they should classify organisms into species and higher taxa. So, much like disputes in the actual metaphysics of race, there are practical payoffs to science or society for engaging in metaphysical disputes in the philosophy of science. However, mainstream analytic metaphysics does not guarantee a payoff for science or society. For instance, what exactly is the payoff for science or society in debating about "the" nature of substantive metaphysics?

Perhaps Sider (2011, 47) sums up my point best when he says, "… this concept is not intended to apply to everything that might justly be called "nonsubstantive". For example, it isn't meant to apply to equivocations between distinct lexical meanings (as in a dispute over whether geese live by "the bank", in which one disputant means river bank and the other means financial bank)… Nor is it meant to capture the shallowness of inquiry into whether the number of electrons in the entire universe is even or odd (an inquiry that is substantive in my sense, but pointless)."

References

Andreasen, R. O. (1998). A New Perspective on the Race Debate. *The British Journal for the Philosophy of Science, 49*(2), 199-225.

Appiah, K. A. (1996). Race, Culture, Identity, Misunderstood Connections. In K. A. Gutmann, *Color Conscious* (pp. 30-105). Princeton: Princeton University Press.

Glasgow, J. (2009). *A Theory of Race.* New York: Routledge.

Glasgow, J., Shulman, J., & Covarrubias, E. (2009). The Ordinary Conception of Race in the United States and Its Relation to Racial Attitudes: A New Approach. *Journal of Cognition and Culture, 9*, 15-38.

Haslanger, S. (2012). *Resisting Reality.* Oxford: Oxford University Press.

Hirsch, E. (2005). Physical-Object Ontology, Verbal Disputes, and Common Sense. *Philosophy and Phenomenological Research, 70*(1), 67-97.

Hochman, A. (2013). Against the New Racial Naturalism. *The Journal of Philosophy, CX*(6), 331-351.

Hochman, A. (2014). Unnaturalised racial naturalism. *Studies in History and Philosophy of Biological and Biomedical Sciences, 46*, 79-87.

Kaplan, J., & Winther, R. (2013). Prisoners of Abstraction? The Theory and Measure of Genetic Variation, and the Very Concept of "Race". *Biological Theory, 7*(1), 401-412.

Ludwig, D. (2015). Against the New Metaphysics of Race. *Philosophy of Science, 82*(2), 244-265.

Maglo, K. (2011). The Case against Biological Realism about Race: From Darwin to the Post-Genomic Era. *Perspectives on Science, 19*(4), 361-390.

Pigliucci, M., & Kaplan, J. (2003). On the Concept of Biological Race and Its Applicability to Humans. *Philosophy of Science, 70*(5), 1161-1172.

Sider, T. (2011). *Writing the Book of the World.* Oxford: Oxford University Press.

Slater, M. (2015). Natural Kindness. *The British Journal for the Philosophy of Science, 66*(2), 375-411.

Spencer, Q. (2014). A Radical Solution to the Race Problem. *Philosophy of Science, 81*(5), 1025-1038.

Templeton, A. R. (2013). Biological Races in Humans. *Studies in History and Philosophy of Biological and Biomedical Sciences, 44*(3), 262-271.

Zack, N. (2002). *Philosophy of Science and Race.* New York: Routledge.

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

Measurement of Statistical Evidence: Picking Up Where Hacking (et al.) Left Off

Abstract Hacking's (1965) Law of Likelihood says – paraphrasing– that data support hypothesis H_1 over hypothesis H_2 whenever the likelihood ratio (LR) for H_1 over H_2 exceeds 1. But Hacking (1972) noted a seemingly fatal flaw in the LR itself: it cannot be interpreted as the degree of "evidential significance" across applications. I agree with Hacking about the problem, but I don't believe the condition is incurable. I argue here that the LR *can* be properly calibrated with respect to the underlying evidence, and I sketch the rudiments of a methodology for so doing.

Introduction

The "likelihoodist," or "evidentialist," school of thought in statistics is well known among philosophers, more so perhaps than among scientists or even statisticians, in large part due to Hacking (1965). One way to distinguish evidentialism from the other major schools – frequentism and Bayesianism – is to note that evidentialism alone focuses on the assessment of statistical evidence as its principal task, rather than decision-making or the rank-ordering of beliefs.[1]

[1] Hacking himself generally prefers the term "support" over "evidence," as does Edwards (1992), but other representatives of this school (Good 1950; Barnard 1949; Royall 1997) refer to an equivalent concept as "evidence." I prefer "evidence," since this is the familiar, albeit vague, word for what we are trying to illuminate; and I prefer "evidentialist" over "likelihoodist" as the name of the school, since the former highlights a key distinction

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

It might be thought, therefore, that evidentialism would be the predominant approach to statistical inference in science, where quantifying evidence is usually the main objective. (If you don't agree, try getting scientists to stop using the p-value as a measure of the strength of the evidence!) But frequentism, and to a lesser extent Bayesianism, predominate in the scientific literature, while evidentialism is virtually unseen. Why is this? I'm going to argue here that the fault lies with evidentialism's failure thus far to address the problem of calibrating the units in which evidence is to be measured. Since meaningful calibration is the sine qua non of scientific measurement, this turns out to be the loose thread that causes the cloth to unravel when we pull on it.

Before proceeding it may be worth noting some things I will and will not be talking about. First, I am concerned only with *statistical* evidence, and will not be considering the concept of evidence as it appears in other contexts, e.g., in legal proceedings. Second, I will treat statistical evidence as a *relationship* between data and hypotheses under a model that can be expressed in the form of a likelihood (as defined below). On this view, data do not possess inherent evidential meaning on their own, but only take on meaning in the context of their relationships to particular hypotheses, with the nature of those relationships governed by the form of the likelihood. I will not be concerned here with measurement problems associated

---

between this school and the others. By contrast, likelihood features prominently in all modern statistical frameworks.

with the data themselves.[2] Third, I am interested here solely in addressing the question of whether this relationship between data and hypotheses can be rigorously quantified. If the answer is yes, then presumably the degree of evidence could play a role in decision making (deciding how strong is strong enough when it comes to evidence) or in guiding belief, but I will not be addressing these topics here. It is one hallmark of evidentialist reasoning that statistical evidence is treated independently of these matters.

The remainder of the paper is organized as follows. In section (1) I articulate the central evidence calibration problem (ECP), and suggest reframing it in measurement terms. In section (2), I consider ways in which evidentialism's preoccupation with so-called "simple" hypotheses (as defined below) has constricted the theory, masking the true nature of the underlying measurement problem, and also obscuring the solution. In section (3) I illustrate a methodology for beginning to address the ECP once the restriction to simple hypotheses is relaxed. In section (4) I briefly consider what changes would be required to axiomatic foundations in order to accommodate this methodology while remaining true to the spirit of evidentialism's original motivating arguments.

(1) The Evidence Calibration Problem (ECP)

At the heart of evidentialism is Hacking's (1965) familiar Law of Likelihood, which says in essence that data support one statistical hypothesis H_1 over another hypothesis H_2

[2] In common usage "evidence" is often used to refer to what I am calling *data*, but "evidence" also has this other sense of being a *relationship* between data and hypotheses. In order to maintain this distinction, I will call the data "data" and the relationship "evidence."

whenever the likelihood ratio (LR) for H_1 over H_2 exceeds 1. But Hacking (1972) pointed out a problem in assigning any particular interpretation to the magnitude of the LR. In his review of Edwards (1992, orig. 1972), he says:

> "Now suppose the actual log-likelihood ratio between the two hypotheses is *r*, and suppose this is also the ratio between two other hypotheses, in a quite different model, with some evidence altogether unrelated to [the original data]. I know of no compelling argument that the ratio *r* 'means the same' in these two contexts."[3] (p. 136)

Thus we can say that, for one experiment, data support hypothesis H_1 over hypothesis H_2 with LR = 2, and, for another experiment, that a different set of data support H_3 over H_4 with LR = 20; but we cannot saying anything definite about how much more the second set of data supports H_3 over H_4 relative to the amount by which the first set supports H_1 over H_2. Edwards was well aware of this problem, saying expressly that "we shall not be attempting to make an absolute comparison of *different* hypotheses on *different* data." (p. 10). But Hacking's point cuts deep. *If the numerical value of the LR cannot be meaningfully compared across applications, in what sense is it meaningful in any one application*?

[3] Here Hacking is using "evidence" in the sense of what I am calling *data*; however, he goes on to describe what he has in mind in terms of levels of "evidential significance." He refers to the *log* LR as this is the form preferred by Edwards. Note that Hacking already appears to have been alluding to this problem in Hacking (1965), vide p. 61.

Hacking's criticism points to a fundamental problem for evidentialists, who appear to be able to say *whether* given data support H_1 over H_2, but not by *how much* they support H_1.[4] This is on the face of it metaphysically perplexing, but also, it leaves a gap between *support*, as Hacking's Law defines it, and a truly quantitative *weight of evidence*, which would be far more useful scientifically if only we could work out how to evaluate it.

Following the core arguments in Barnard (1949), Hacking (1965) and Edwards (1992), I will assume that the LR is the key quantity in any cogent theory of statistical evidence. But the Law of Likelihood is more specific than this assumption: it assigns a particular importance to one very narrowly conceived *aspect* of the LR, a fact that is obscured by evidentialism's focus on simple hypotheses, to which I turn next.

Before doing so, I note that resolving Hacking's problem requires unpacking his phrase 'means the same'. I think that this must be understood as 'means the same with respect to the underlying evidence,' a locution that lands us solidly in *measurement* territory. We must be able to think in terms of the underlying evidence, as something we can – at least in the abstract – conceive of independently of how we measure it. The question then becomes: How do we establish meaningful measurement units for evidence, so that a given measurement value always 'means the same' *with respect to the evidence*? This is the ECP.

And here, in a nutshell, is the evidentialist's difficulty in addressing the ECP. The LR for a simple hypothesis comparison (see below) is a single number, thus, the evidentialist is lured

[4] Royall (1997) is the only one as far as I know who argues that the magnitude of the LR *does* express strength of evidence in a comparable manner across applications. But I think his arguments on this point fail for reasons articulated in Forster & Sober (2004).

into the claim that "the LR *is* the evidence." To see the danger here, consider a mercury thermometer reading 80°F. We might say, "the temperature is 80°," but this is a circumlocution for "80 is the numerical value we assign, on the Fahrenheit scale, to the underlying temperature." Now suppose that rather than degrees, only units of volume *V* are annotated on the sides of the glass. We might be tempted to say "*V* is the temperature," but now this statement is not merely a circumlocution, it is also an error. *V* alone does not tell us the temperature; we must, at the least, also take into account the pressure. To insist that temperature can be represented by volume alone, or by pressure alone, or by any other single thing that can be readily and directly measured, is to mistake the nature of temperature. Just so, I am going to argue that *the simple LR mistakes the nature of evidence*, by obscuring the fact that the evidence itself is not a number, and moreover, that the evidence is not any single thing that can be readily and directly measured, but instead, it is a function of (at least) two measurable things.

(2) The Insidiousness of Simple Hypotheses

To begin with, we need to define *likelihood*:

> "The likelihood, L(H|R), of the hypothesis H given data R, and a specific model, is proportional to P(R|H), the constant of proportionality being arbitrary." Edwards (1992) (p. 9)

Two key points are familiar: (i) likelihood represents a feature of an hypothesis given data, not the other way around; and (ii) likelihood is related to but not the same as probability,

since it is defined only up to an arbitrary multiplicative factor and therefore does not follow the Kolmogorov axioms. I will not rehearse the advantages of likelihood in spelling out a theory of statistical evidence, but suffice it to say that likelihood enables inferences to proceed independently of what are, arguably, extraneous features of study design, including the sampling distribution of all those observations that might have occurred but didn't.

There is a third important feature of this definition as well, and this regards the nature of the *hypotheses* to which the definition is intended to apply. Edwards is, as always, explicit:

> "An essential feature of a statistical hypothesis is that its consequences may be described by an exhaustive set of mutually-exclusive outcomes, to each of which a definite probability is attached." (p. 4)

This precludes consideration of likelihoods involving *composite* hypotheses. For instance, in the context of a coin-tossing experiment in which x independent tosses have landed heads and y have landed tails, and letting θ=P(heads), one can write the likelihood $L(\theta=0.1|x, y)$, or $L(\theta=0.2|x, y)$. These likelihoods involve "simple" hypotheses, in which θ is assigned a single numerical value, so that the corresponding probability $P(x, y|\theta)$ returns a single number on the probability scale for each possible outcome (x, y). But one can *not* write $L(\theta=0.1$ or $\theta=0.2|x, y)$, because the latter involves a "composite" hypothesis, which does not assign a definite probability to the observed outcome. To know the probability of observing (x, y) under the hypothesis "θ=0.1 or θ=0.2," we would need not only to know the probability of (x, y) for each θ, but also, we would need to know the prior probabilities of θ=0.1 and θ=0.2. As these prior probabilities lies outside the likelihood, they are not admissible on the

evidentialist view.

But even the simplest examples of statistical reasoning generally involve hypotheses that appear on the face of things to be composite; e.g., we might be interested in whether the coin is biased toward tails or fair, which would appear to involve the improperly formed hypothesis $\theta<0.5$. This situation is handled by treating composite hypotheses "solely on the merits of their component parts" (Edwards, p. 5). Thus in forming the LR corresponding to 'coin is biased toward tails' vs. 'coin is fair,' we would need to consider separately the (infinitely many) simple LRs in the form $L(\theta=\theta_i|x, y)/L(\theta=0.5|x, y)$, for each possible i^{th} value of $\theta\leq 0.5$. Now the LR is a function of θ, not a single number (Figure 1).

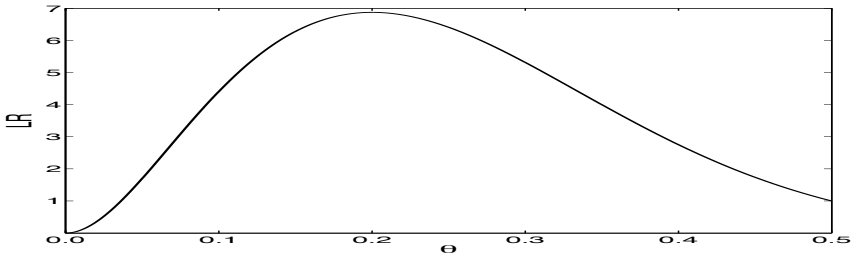


Figure 1 LR as a function of θ for $x = 2$, $y = 8$.

In practice it seems that what is important is not so much the proscription against composite hypotheses, but rather the prescription for how they may be interpreted. We can graph the LR as a function of θ, as if we were admitting composite hypotheses, but we can only make statements like "$\theta=0.2$ is supported over $\theta=0.5$, on given data, by LR=6.9," while

"θ=0.1 is supported over θ=0.5, on those same data, by LR=4.4."[5] But as a practical matter, the graph is not a sufficiently concise summary for general scientific applications. We still need some way to reduce the function LR(θ) to a single number summarizing the strength of the evidence.

And this is where we get into trouble, because focus shifts naturally to the *maximum* LR (MLR), which occurs over the best supported value – the maximum likelihood estimate (m.l.e.) – of θ. Indeed, given that we are only allowed to make statements about one simple hypothesis comparison at a time, the MLR, itself a ratio of two simple likelihoods, appears as the best single constituent LR to use as a summary feature of the LR graph. (Below I consider how relaxing the requirement that hypotheses must be simple frees us up to consider other features.) We have now successfully summarized the *function* LR(θ) as a single number, the MLR, but this summary is tethered to the m.l.e.. We appear to have answered the question: How well supported is the m.l.e. compared to (one or more individual) alternative values of θ? But that is not the question we asked initially, which was about the evidence.[6]

The m.l.e. of θ arrives on the scene as a seemingly innocuous point of special interest, the value that corresponds to the maximum support, but it rapidly takes over, embroiling us in a downward spiral of increasingly perplexing difficulties. One immediate issue with relying on the MLR to summarize the evidence (continuing to focus for ease of discussion on the coin-

[5] Moreover we can only make such statements when both the data and the form of the likelihood are the same in the numerator and the denominator of the LR, for only in such cases will the constants of proportionality cancel.
[6] Hacking (p. 28 ff.) makes clear the conceptual reasons for keeping estimation and evidence (or support) separate.

tossing example, in which maximization occurs only in the numerator of the LR), is that MLR≥1: the MLR can only show evidence in favor of the numerator but never in favor of the denominator. This is problematic, like using a thermometer in which the mercury is prevented from receding.

Another problem with the MLR is that it begs the question of measurement scale in a particularly obvious way, because its evidential meaning would appear to require some kind of adjustment to compensate for the maximization itself. The more parameters we maximize over (again, for ease of discussion, assuming maximization occurs only in the numerator), the larger the MLR becomes. How are we to separate the portion of the MLR reflecting the evidence from the portion representing an artifact of the process of maximization? It becomes particularly hard to retain the fiction that the numerical value of the *maximum* LR has some prima facie meaning with respect to the underlying evidence, regardless of the number of parameters over which the LR is maximized.

There is a third, more subtle but at least as damaging, difficulty with summarizing evidence via MLRs. Simple LRs can be multiplied across two data sets, but MLRs can not be multiplied. Rather, to obtain the MLR based on two sets of data, we first combine the data to find the new m.l.e., which is a kind of weighted average of the two original m.l.e.s, and then we find the new MLR with respect to this average m.l.e. on the combined data. Now consider a situation in which data set D_1 favors H_2 by some substantial amount, and D_2 also favors H_2, but by a lesser amount. In such situations it is not uncommon for the combined support for H_2 to be less than the original support on D_1 alone. But this is not how *evidence* behaves:

strong evidence for H_2 followed by weaker evidence also supporting H_2 ought to lead to *stronger* evidence for H_2, not intermediate evidence. (A blood type match following a DNA match does not lessen the evidence that the defendant was at the crime scene.[7]) This means that we cannot in practice differentiate between situations in which new data are truly diminishing the evidence, and situations in which the evidence is in fact increasing but the MLR at the average m.l.e. goes down anyway. This tendency of the MLR to "average" across combined data is entirely due to its dependence on the m.l.e.; simple LRs do not share this defect.[8]

Of course none of this need surprise unreconstructed evidentialists, who, after all, disavowed composite hypotheses – and therefore any need for maximization – from the start. But then beyond the simplest of examples, we are left with an irreducible graph of the component simple LRs, not a single number. This is true already in single-parameter cases; the problem is only exacerbated in higher dimensions.

There is also the matter of masking the nature of the real problem: by focusing initially only on those situations in which the LR is a single number, we missed Hacking's *measurement* question, how do we ensure that this number always 'means the same'? It is only when we consider composite hypotheses that it becomes clear we were never warranted

[7] This example was suggested by Hasok Chang.
[8] This issue plays a salient role in the current "crisis" of non-replication of statistical findings in the biomedical and social sciences, where the tendency of p-values and MLRs to "regress to the mean" upon attempts to replicate initial findings is widely interpreted as meaning that the evidence has gone down. In the absence of a properly behaved evidence measure, however, this conclusion is entirely unwarranted.

in the first place in assuming that the face value of the LR for a simple vs. simple hypothesis comparison *is* the evidence. Composite hypotheses force us to think in terms of the LR graph, which, precisely because it is not a single number, immediately raises the issue of which *feature(s)* of the graph might be relevant to the evidence. Composite hypotheses are crucial, not only because they are scientifically relevant, but also, because they beg a question all but hidden as long as we focus only on simple hypotheses.

The urge to sidestep the problem of the evidential interpretation of the MLR is the reason evidentialists have been reluctant to admit composite hypotheses into their formalism in the first place. But it is fair to say that they have failed to provide any viable alternative to the MLR as the summary measure of evidence strength in practice. The preoccupation with simple hypotheses has entailed inherent difficulties for the program, and it has also masked a basic underlying calibration issue. The good news, I believe, is that it has also been masking the possibility of a solution.

(3) Towards a Solution to the Measurement Calibration Problem

Consider again the coin-tossing experiment and LR(θ) as shown in Figure 1. Let us suppose, following the spirit if not the letter of the Law of Likelihood, that all of the evidential information is captured, somehow, in this graph. What *feature(s)* of the graph should we take as representing the degree of evidence?

The MLR of course is one possibility, but I have already stated some objections to this option. An alternative would be to use the *area* under the graph (ALR). (Note that this is

only possible if we allow ourselves to consider the truly composite hypothesis $\theta<0.5$, because the ALR requires simultaneous consideration of all of the constituent simple hypotheses.[9]) But while we're at it, why not also consider using *sets of features* of the graph? For instance, the evidence might be a function of the both the MLR and the ALR, e.g., their product, or their ratio. What we need is a methodology for figuring out which among the many possibilities is the correct one.

The methodology I propose is quite simple, at least to begin with. Let's consider the *behavior of candidate evidence measures* in situations where we have clear intuitions regarding the *behavior of evidence*, and see which of our candidate measures behaves like the object of measurement, the evidence. Here I will illustrate using coin-tossing "thought experiments" to discover patterns of behavior of the evidence with changes in data, considering the evidence that the coin is either biased toward tails or fair. I propose that, perhaps with a little persuasion, I could convince you that the following patterns capture *what we mean* when we talk about statistical evidence in this context. (Here I summarize the data in terms of n=the number of tosses, and x/n=the proportion of tosses that land heads.)

(i) Evidence as a function of changes in n for fixed x/n For any given value of x/n, the evidence increases as n increases. The evidence may favor bias (e.g., if $x/n = 0.05$) or no bias (e.g., if $x/n = ½$), but in either case it gets stronger with increasing n.

[9] The ALR is proportional in this simple example to the Bayes factor under a uniform prior on θ, which is sometimes interpreted in Bayesian circles as a measure of evidence strength; it is also proportional to the relative belief (Evans 2015), another Bayesian proposal for measuring evidence. But the ALR itself does not involve a prior, so I see no prima facie reason for the evidentialist to balk at this suggestion, once composite hypotheses are allowed.

(ii) Evidence as a function of changes in x/n for fixed n If we hold n constant but allow x/n to increase from 0 up to, say, 0.20, the evidence favoring 'coin is biased' diminishes: i.e., the evidence for bias is stronger the further x/n is from ½. But we have also already noted that when x/n is close to ½ the evidence favors 'coin is fair.' Therefore, as x/n continues to approach ½, at some point the evidence will shift to favoring 'coin is fair,' and from that point, the evidence for 'coin is fair' will increase the closer x/n is to ½.

(iii) Rate of evidence change as a function of changes in n for fixed x/n For given x/n, as n increases the evidence *increases more slowly* with fixed increments of data. E.g., consider evidence in favor of bias with one additional tail (T), following T, or TT, or TTT. When the number of tails in a row is small (i.e., when there is weak evidence favoring bias), each subsequent T makes us that much more suspicious that the coin is biased. But suppose we have already observed 100 Ts in a row: now one additional T changes our sense of the evidence hardly at all, as we are already quite positive that the coin is not fair.[10]

(iv) x/n as a function of changes in n (or vice versa) for fixed evidence It follows from (i) and (ii) that in order for the *evidence* to remain constant, n and x/n must adjust to one another in a compensatory manner. E.g., if x/n increases from 0 to 0.05, in order for the evidence to remain the same n must increase to compensate; otherwise, the evidence would go down, following (ii) above. By the same token, it is readily verified that if (i)

[10] This underscores the point made above that evidence is not inherent in the data (say, a single toss T), but rather, evidence is a relationship between the data and the hypotheses that depends on context.

and (ii) hold, then as x/n continues to increase, at some point n must begin to decrease in order to hold the evidence constant as the evidence shifts to favoring 'coin is fair.'

Note that at this point we have not mentioned probability distributions, likelihoods, or parameterization of the hypotheses. These patterns characterize evidence in only a very informal, vague manner. However, by the same token, they exhibit a kind of generality: they derive from our general sense of evidence, from what we *mean* by statistical evidence before we attempt a formal mathematical treatment of the concept.

Can we find a precise mathematical expression that exhibits these patterns? As illustrated in Figure 2, the *ratio RLR=MLR/ALR exhibits all of the expected behaviors*. By contrast, neither MLR nor ALR shows all four of these patterns. For instance, MLR, as already noted, cannot show increasing evidence in favor of H_2 because it can never favor H_2 in the first place; and both MLR and ALR increase exponentially in n for fixed x/n rather than showing the concave-down pattern in 2(a).

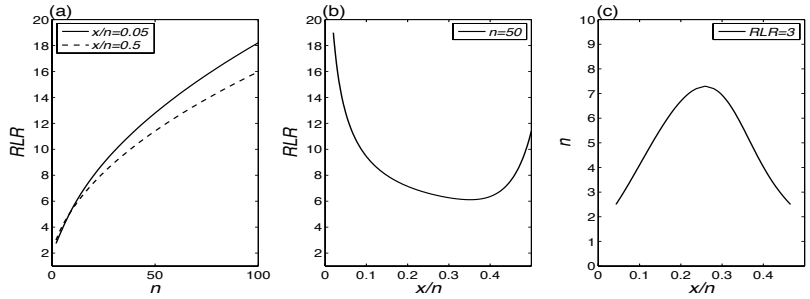


Figure 2 Patterns of behavior of RLR for coin-tossing thought experiments: (a) Patterns (i) and (iii); (b) Pattern (ii); (c) Pattern (iv).

Of course none of this proves that RLR is the correct, or optimal (or properly calibrated) measure of evidence. But this style of reasoning buys us an important methodological tool. Whichever features of the LR graph we consider and however we combine them, we must be able to show that the resulting evidence measure *behaves like the evidence*. When proposing candidate evidence measures anything goes, but only those candidates that behave appropriately remain on the ballot. And even in this very simple example, two obvious candidates – the MLR and the ALR – have already dropped out of contention.

Of course, there is no reason to assume that what works in this simple case (RLR) will work in more complicated cases, nor have we yet resolved the ECP's fundamental calibration issue. Establishing that a measure behaves like the object of measurement is only a first step, but it is a vital step not previously taken. It provides an "empirical" measurement scale, not an absolute scale, much as early thermoscopes provided good experimental tools while falling short of proper, absolute, calibration (Chang 2004).[11] Projecting an empirical measure onto an absolute scale requires a broader theoretical foundation, but one needs the empirical measure first. My point here is simply that confronting the ECP head on, and in the context of composite hypotheses, opens the door for the first time to the possibility of establishing a proper measurement scale for statistical evidence.

Note too that the coin-tossing exercise suggests the existence of an *equation of state* involving the three quantities (n, x/n and the evidence), such that fixing any one quantity

[11] Indeed, the ECP poses what Chang calls a "nomic" measurement problem, much like the nomic problem of temperature measurement. What I am describing here is a necessary but not sufficient stage in resolving a nomic problem.

while allowing a second one to change requires a specific compensatory change in the third. This in turn suggests a new, and potentially very powerful, way to think about the laws governing the behavior of LRs. I'm not aware of any evidentialist work that considers such equations, but I see no reason that an evidentialist-at-heart should be prohibited from pursuing their study.

(4) Relaxing the Foundations To Include Composite Hypotheses

In order to tackle the ECP in the terms of the preceding section, we need to amend the foundations of evidentialism, but only slightly. I propose the following changes. First, let's retain Edwards definition of likelihood, as quoted above, but insert the word "simple" (which is tacit in Edwards' original statement): "The likelihood, L(H|R), of a *simple* hypothesis H given data R, and a specific model, is proportional to P(R|H), the constant of proportionality being arbitrary." Second, we can again add the word "simple" to his characterization of a statistical hypothesis: "An essential feature of a *simple* statistical hypothesis is that its consequences may be described by an exhaustive set of mutually-exclusive outcomes, to each of which a definite probability is attached." But we can now add a definition of likelihood for a composite hypothesis: "A *composite* hypothesis H given data R, and a specific model, is the set of all constituent simple hypotheses, defined up to a single constant of proportionality." Thus the essential feature of a *composite* hypothesis is that *each of its constituent simple hypotheses* may be described by an exhaustive set of mutually-exclusive outcomes, to each of which a definite probability is attached. We can now use this definition

of a composite hypothesis to define the corresponding composite likelihood, as the set of all constituent simple likelihoods.

Under my proposal, the spirit of the Law of Likelihood can be retained: We can say that all of the *evidential information* conveyed by given data regarding a comparison between two hypotheses on a particular model is contained in the LR, where, under the expanded definition of hypotheses, the LR is understood to be a function of all unknown parameters, or better still perhaps, a *graph*. This can equivalently be read as a definition of *evidential information,* as whatever changes the LR graph.[12] But the idea that the (simple) LR itself expresses the degree or weight of the evidence must be abandoned. What I have attempted to argue here is that there is at least the possibility of replacing this notion with something more useful.

Discussion

Evidence is a general and vague term in science. Statistical evidence is a narrower concept, but it still inherits some of this vagueness. One way to tackle a general and vague term is by seeking a precise definition that maintains full generality, but of course, this might not be possible. Weyl (1952) has suggested another approach:

> "To a certain degree this scheme is typical for all theoretic knowledge: We begin with some general but vague principle, then find an important case where we can give that

[12] I borrow this idea from Frank (2014), who defines *information* as whatever changes a probability distribution.

> notion a concrete precise meaning, and from that case we gradually rise again to generality… and if we are lucky we end up with an idea no less universal than the one from which we started. Gone may be much of its emotional appeal, but it has the same or even greater unifying power in the realm of thought and is exact instead of vague." (p. 6)

Can evidentialism be redeemed and made truly useful to science? Of course I have not proved that the answer is yes. But in section (3) I illustrated a case in which we appear to be able to give the vague concept of statistical evidence a concrete, precise meaning, via the quantity RLR=MLR/ALR. It remains to be seen whether it is possible to rise again to generality from this first step. But for those of us who agree with most of what Barnard, Hacking and Edwards have to say on the subject, it seems worthwhile to see how far we can take this line of reasoning. This also seems to be a singular opportunity for philosophers of science to step into the breach and at least *try* to solve a problem that has long stood between one of the needs of science – for well-behaved quantitative measures of evidence – and the capabilities of conventional statistical methodologies.

References

Barnard G.A. "Statistical Inference." *J Royal Stat Soc* XI, no. 2 (1949):115-39.

Chang H. *Inventing Temperature: Measurement and Scientific Progress*. New York:Oxford UP, 2004.

Edwards A.W.F. *Likelihood*. Baltimore:Johns Hopkins UP, 1992. Orig. Cambridge UP, 1972.

Evans M. *Measuring Statistical Evidence Using Relative Belief*, Monographs on Statistics and Applied Probability. Boca Raton:CRC Press, Taylor & Francis Group, 2015.

Forster M, Sober E. "Why Likelihood?" In *The Nature of Scientific Evidence*, Taper & Lele eds., 153-90. Chicago:Chicago UP, 2004.

Frank S.A. "How to Read Probability Distributions as Statements About Process." *Entropy* 16(2014):6059-98.

Good I. J. *Probability and Weighing of Evidence*. London:Griffon, 1950.

Hacking I. *Logic of Statistical Inference*. London:Cambridge UP, 1965.
———. "Review of Edwards' Likelihood." *British J Phil of Sci* 23(1972): 132-37.
Royall R. *Statistical Evidence: A Likelihood Paradigm*. London:Chapman & Hall, 1997.
Weyl, Hermann. *Symmetry*. Princeton UP, 1952.

What Basic Emotions Really Are

Encapsulated or Integrated?

Abstract: While there is ongoing debate about the existence of basic emotions (BEs) and about their status as natural kinds, these debates usually carry on under the assumption that BEs are encapsulated from cognition and that this is one of the criteria that separates the products of evolution from the products of culture and experience. I aim to show that this assumption is entirely unwarranted, that there is empirical evidence against it, and that evolutionary theory itself should not lead us to expect that cognitive encapsulation marks the distinction between basic and higher cognitive emotions. Finally, I draw out the implications of these claims for debates about the existence of basic emotions in humans.

1. Introduction

It is widely held among emotion theorists that there is some theoretically interesting distinction between basic and higher cognitive emotions. On this picture, basic emotions (BEs) are primarily structured by evolution whereas higher cognitive emotions are substantially structured by either culture or individual experience. While there is ongoing debate about the existence of BEs and about their status as natural kinds, these debates usually carry on under the assumption that BEs are encapsulated from cognition and that encapsulation is one of the criteria that separates the products of evolution from the products of culture and experience. I aim to show that this assumptions is entirely unwarranted, that there is empirical evidence against it, and that evolutionary theory itself should not lead us to

Isaac Wiegman
10/19/2016

expect that cognitive encapsulation marks the distinction between basic and higher cognitive emotions. Finally, I draw out the implications of these claims for the existence of basic emotions in humans.

In the following section, I characterize the received view of BEs, which holds (among other things) that BEs are solutions to *basic life problems* in our evolutionary past. Then I consider and reject some of the reasons to think that BEs are cognitively encapsulated. In the second section, I provide an example of a BE in rodents that bears the marks of cognitive integration (as opposed to encapsulation). The basic life problem that likely shaped this emotion appears to demand substantial cognitive integration. In the third section, I draw out the implications for a current debate in emotion theory concerning the existence of BEs in humans.

2. Basic Emotions

BEs – including anger, fear, happiness, sadness, disgust, and surprise (for an extended list, see Ekman & Cordaro, 2011) – are thought to be human-typical behavioral syndromes that include involuntary facial expressions of emotion, physiological changes (e.g. in heart rate, blood pressure, and hormone levels), and changes in bodily posture (including bodily social displays and orienting responses). According to BE theory, these syndromes have a similar kind of evolutionary explanation and similar neural and psychological mechanisms. Specifically, they each evolved to address basic life problems or adaptive problems (such as

Isaac Wiegman
10/19/2016

resource competition, avoidance of predators and avoidance of poisons and parasites). Some of these basic life problems are ones that we share with non-human animals.

Moreover, the elicitation and production of these syndromes (including the coordination of various response components) are supposed to be explained by *automatic appraisal mechanisms* and *affect programs*, respectively (Ekman, 1977, 1999). For instance, affect programs explain phenomena observed in experiments that ask people to distinguish photographs of facial expressions of emotions, connect these expressions with emotion terms, or rate their appropriateness in response to vignettes (for an overview, see Ekman, 2003). They are also supposed to explain the results of experiments that connect facial expressions with changes in physiological response components (Ekman, Levenson, & Friesen, 1983; Levenson, Ekman, & Friesen, 1990). To generalize, affect programs are introduced to explain the observed coordination of various response components and the cross-cultural production of these various syndromes (which is thought to explain widespread recognition of facial expressions across cultures).

3. Unwarranted Assumptions Concerning Cognitive Integration

Many emotion theorists claim that BEs lack cognitive integration. In this section, I argue that these claims are based on unwarranted assumptions.

Assumption 1: Cognitively Integrated only if Informationally Integrated

In most cases, questions about the integration of emotions with cognition concern the possibility that emotions are modular in Fodor's (1983) sense. This depends (among other

Isaac Wiegman
10/19/2016

things) on whether they can store *information* that cognitive systems cannot access (*informational encapsulation*); or whether *information* from other cognitive systems can interfere with the operations of an emotion (*cognitive penetrability*); or whether people have conscious access to emotional processes or merely their outputs (*opacity*); or whether the *information* that an emotion provides is general as opposed to specific (which would imply *shallow outputs*). These are some of the more well-known marks of cognitive integration or its absence, encapsulation.

Philosophers and psychologists alike usually proceed under the assumption that integration with cognition depends entirely on whether information is integrated in these ways. These assumptions translate to discussions about BEs, where evidence for lack of *informational* integration is sometimes used as evidence for lack of *cognitive* integration *simpliciter*:

> Three other types of evidence suggest that [basic] emotion processes can operate independently of cognition. Emotions have been induced by unanticipated pain…, manipulation of facial expressions…, and changing the temperature of cerebral blood… In all these conditions the immediate cause of the emotion was noncognitive. (Izard, 1992, p. 563, see also his 2007)

Here, Izard apparently assumes that the impenetrability of BEs constitutes evidence that BEs operate independently of cognition. The fact that they respond to low level inputs or processes to which other systems have limited access certainly suggests that emotional states can respond to information that is not integrated with cognition. In addition, there is evidence

Isaac Wiegman
10/19/2016

that people cannot fully control facial expressions of BEs (Ekman, 1972; Friesen, 1973), suggesting that BEs are cognitively impenetrable. Overall, BEs appear to lack informational integration.

Nevertheless, the realm of the cognitive picks out not only informational states, but also includes a broader range of internal states that function as causal intermediates between stimulus and response, perception and action (Rey, 1997). Cognitive states so understood include not only informational states (such as beliefs) but also motivational states (such as desires). Moreover, questions about cognitive integration may be asked about either informational or motivational states. If so, the possibility arises that the two forms of cognitive integration are independent of one another. If so, any inference from the one to the other is invalid.

This becomes clear when we consider hunger. Hunger may very well be akin to desire (a paradigmatic case of a cognitively integrated state) in the sense that it can interact with other cognitive systems to produce flexible or novel behaviors, as when rodents take novel "short cuts" to get to a food box in a maze (Olton, 1979; Tolman, 1948). Short cut behaviors suggest that hunger is a motivational state that can incline rodents to the pursuit of an end (e.g. food consumption) by selecting from a range of different means, perhaps by interacting with informational states that relate means to ends (e.g. means-ends beliefs). Even so, hunger may be cognitively impenetrable in that it may be triggered by low level stimuli and processes (e.g. low-level detection of changes in blood sugar). Moreover, when one feels hungry, one cannot interfere with the feeling of hunger by thinking about it (e.g. by noticing

Isaac Wiegman
10/19/2016

that the amount of energy one's body has stored in fat deposits is more than enough to sustain oneself). One can even imagine that it is informationally encapsulated: it might store information (e.g. about which foods are more calorically dense) that other systems cannot directly access.

These conceptual possibilities suggest that questions concerning the integration of informational states are conceptually independent of questions concerning the integration of motivational states. Hunger may be informationally encapsulated while retaining a degree of integration as a motivational state. Wholesale encapsulation, therefore, does not follow from informational encapsulation. If this is correct, then inferences like the one Izard draws above are invalid: having non-cognitive inputs is not a reason to think that emotions operate independently of cognition. They might very well operate in concert with cognition on the output side or as motivational states. Before I raise that possibility, consider another reason to rule it out at the outset: that BEs are not integrated with propositional attitudes, including beliefs *and* desires.

Assumption 2: Integration with Beliefs and Desires is the Criterion for Cognitive Integration

Contrary to the previous assumption, this one respects the distinction between motivational and informational integration. Nevertheless, I argue that it sets the bar for cognitive integration too high.

Isaac Wiegman
10/19/2016

To see this, consider Griffiths' (Griffiths, 1997, 2004) views on the distinction between basic and higher cognitive emotions. First, he draws on some of the same evidence as Izard to conclude that BEs are opaque and informationally encapsulated. Since they have these and other marks of modularity, Griffiths thinks BEs have "limited involvement" with higher cognitive processes, which are "…the processes in which people use the information of the sort they verbally assent to (traditional beliefs) and the goals they can be brought to recognize (traditional desires) to guide relatively long-term action and to solve theoretical problems." (Griffiths, 1997, p. 92) Here, Griffiths may be making the same faulty assumption as Izard (that informational encapsulation implies cognitive encapsulation more broadly). However, let us grant that he may have additional reasons to think that emotions are not integrated on the output side or qua motivational states.

From this, Griffiths draws a broader conclusion: that BEs are not "flexible [or] integrated with long-term, planned action" and are instead "restricted to short-term, stereotyped responses" (Griffiths, 1997, p. 241). The apparent assumption is that if BEs are not integrated with beliefs, desires and long-term planning, then the only alternative is that they are similar to fixed action patterns, being inflexible and stereotyped. Griffiths makes no explicit argument for this assumption, perhaps at the time it was widespread enough to make further argument otiose.

Nevertheless, it has become a tendentious assumption for several reasons. First, the phenomena of intelligent action are much broader than deliberate, "long-term, planned action" mediated by beliefs and desires. For instance, Ginet (1990) argues that many clear

Isaac Wiegman
10/19/2016

cases of actions (as distinct from mere behaviors, such as reflexes or fixed action patterns) are not plausibly mediated by conscious beliefs, desires or intentions: involuntarily crossing one's legs, kicking a door in anger, impulsively pulling a loose thread from one's clothes, and slamming on the brakes to avoid hitting a dog. These actions are not mere behaviors or reflexes. That is, they appear to be purposive and guided by the agent, but it is difficult to find belief-desire style explanations that render them intelligible.[1] Why not think that BEs can influence actions more akin to this variety than to "long-term, planned actions"? Griffiths never raises this question, neither does he give reason to rule out the possibility that BEs cause actions intermediate between long-term planned action and stereotyped behavioral responses.

Second, if we ask what might explain the other varieties of action that Ginet picks out, it may be that such actions are guided by other representational states, aside from conscious or verbally reportable beliefs, desires and intentions. For instance, in the last twenty years, cognitive scientists have begun to emphasize the role of unconscious or non-conceptual representational states in generating flexible and intelligent behavior (Bermúdez, 2003). Informational states aside from beliefs include perceptual representations, map-like spatial representations and representations of affordances. Motivational states aside from desires include drives, incentives and feedback mechanisms.

[1] See also Hursthouse (1991).

Isaac Wiegman
10/19/2016

The flexibility and intelligence of these representational states becomes clear when we consider animal behavior. Nonhuman animals display forms of intelligent or purposive or instrumental behavior (see e.g. Balleine & Dickinson, 1998), even while lacking linguistically mediated propositional attitudes. This suggests that instrumental behaviors in non-human animals are underwritten by a different form of cognitive integration. Consider what Susan Hurley calls *holistic flexibility*:

> The holistic flexibility of intentional agency contributes a degree of generality to the agent's skills: a given means can be transferred to a novel end, or a novel means adopted toward a given end. The end or goal functions as an intervening variable that organizes varying inputs and outputs and allows a degree of transfer across contexts. (Hurley, 2003, pp. 237–38)

Where this sort of flexibility is found, it suggests that behavior is best explained with reference to informational states which represent the means available to an organism (e.g. affordances) and motivational states that represent its ends (e.g. drive states), which can interact interchangeably in order to bring about the same end by various means or to deploy a single means to bring about various ends.

Nevertheless, these informational and motivational states may sometimes lack inferential integration with beliefs and desires. Even in humans, phenomena like "blind-sight" suggest that perceptual representations can flexibly guide behavior without being integrated with verbally reportable states. That is, even though these perceptual states are not verbally reportable or consciously accessible, these informational states mediate goal-

Isaac Wiegman
10/19/2016

directed behaviors (e.g. putting a plate in a slot) rather than just reflexes and fixed action patterns (see e.g. Goodale, Milner, Jakobson, & Carey, 1991). All this suggests that Griffiths' requirements on cognitive integration are too stringent. Verbal reportability and conscious accessibility of a representational state is not necessary for such a state to influence flexible behaviors. To my knowledge there is no evidence that BEs fail to meet less stringent requirements on cognitive integration such as holistic integration.

Once the full range of representational states is expanded in this way (beyond beliefs and desires), it becomes possible that BEs have some degree of motivational integration with other representational states aside from conscious beliefs and desires to produce behaviors that are more flexible and purposive than stereotyped behaviors. Griffiths provides no reason to rule out this possibility.

4. Evidence of Integration in a Basic Emotion

In fact, there is some reason to rule it in. Consider the instinctive patterns of territorial behavior of rodents. These behaviors have been investigated in great detail using a resident-intruder experimental paradigm (for an overview, see D. C. Blanchard & Blanchard, 1984, 2003) add it Adams RRR) in which resident (who have occupied a cage or colony for a few weeks) will attack unfamiliar male intruders introduced into their cage. The attacks of the resident and the defensive maneuvers of the intruder comprise sets of stereotyped behaviors. Each attack behavior of the resident is paired with a matching defensive maneuver of the intruder. The resident adopts a set of stereotyped postures and attacks aimed at biting the

Isaac Wiegman
10/19/2016

dorsal surfaces of the intruder. On the other hand, the intruder adopts a distinctive set of stereotyped behaviors aimed at avoiding or blocking the resident's attempts to bite its back.

While these behaviors are certainly stereotyped, they are not brittle or reflexive. For instance, attacks of residents vary depending on the defensive strategy adopted by the intruder, and they seem to be governed by a motive to approach and attack that persists the entire time that the intruder is present. By contrast, the intruder rat's whole suite of behaviors seems to be governed by a persistent motive to escape and avoid.

Isaac Wiegman
10/19/2016

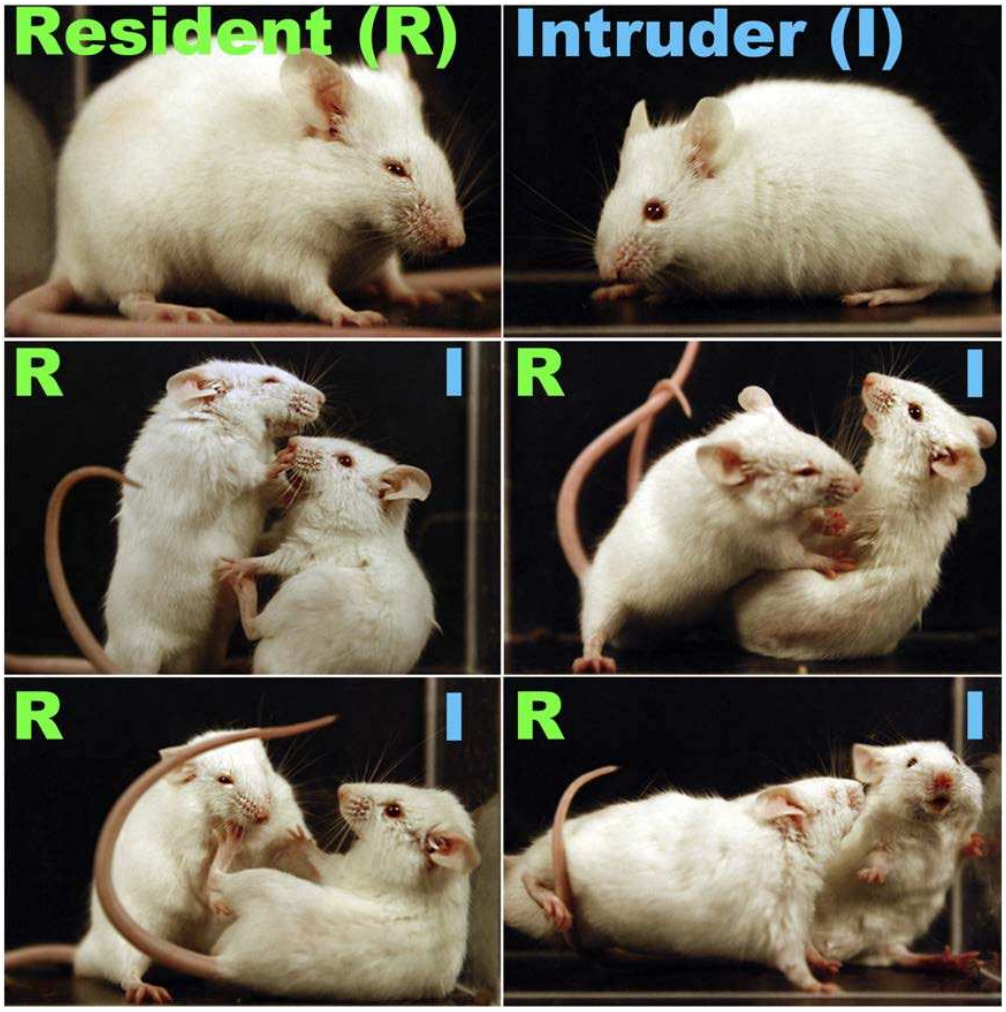


Figure 1 Confrontation and avoidance behaviors (e.g. facial expressions, postures and maneuvers) of resident and intruder mice (respectively). From Defensor and Corley (2012), p. 683 permission pending © Elsevier. Originally published in Physiology and Behavior.

Isaac Wiegman
10/19/2016

What scientists have discovered about these behaviors (the flexibility of these behaviors and their coherent aims) indicates that they are produced by two underlying motivational systems, what I call the confrontation and avoidance systems (D. C. Blanchard & Blanchard, 1984, 2003; D. C. Blanchard, Litvin, Pentkowski, & Blanchard, 2009). The confrontation system is tuned to bring about a specific end state, repeated back-biting. Moreover, this motive does not depend on learning: rats which have been socially isolated from birth will still attempt to bite the back of an intruder (Eibl-Eibesfeldt, 1961). So far, the focus has been on cases in which a given rodent is purely motivated by confrontation or avoidance, but aggressive encounters in the wild usually involve a mix of offensive and defensive postures. This suggests that these motivational systems can be activated simultaneously or in close succession to produce mixed patterns of behavior.

Regardless, these systems have many of the characteristics of affect programs in humans. They are posited to explain a coordinated suite of behaviors and physiological changes that may include facial expressions, cardiovascular changes, and endocrine responses (Defensor, Corley, Blanchard, & Blanchard, 2012; Fokkema, Koolhaas, & van der Gugten, 1995). Moreover, these systems are tailored to solve basic life problems. Specifically, the confrontation system solves the problem of defending territories from other males for breeding purposes (and without fatally injuring kin in the process), whereas the avoidance system solves the problem of avoiding occupied territories and failing that, defending against the attacks of residents. For these reasons, we have all the same reasons to

Isaac Wiegman
10/19/2016

postulate BEs in rodent that we have in humans. Let us suppose then that the confrontation and avoidance systems are BEs in rodents.

Interesting for my purposes, under certain conditions, the presence of the unfamiliar male can produce highly flexible and novel behaviors. In the bound-intruder task, an intruder is tied down on a Plexiglas plate with only its ventral surfaces (belly-side) exposed and placed in the cage of a resident, so that the resident cannot easily bite the back of the intruder. As a result, the resident will sometimes bite at the bands that tie down the intruder or dig under the intruder so that the resident can bite the intruder's back (R. J. Blanchard, Blanchard, Takahashi, & Kelley, 1977). In contrast, none of these behaviors are adopted when the intruder is tied down with his back exposed.

These instrumental behaviors are clearly not stereotyped forms of attack, rather they are forms of flexible behavior adjustment to achieve the aim of biting the intruder's back: they exhibit holistic integration. In this case, the same end can be achieved by several, novel means. Attempts to bite the intruder's bonds or to dig underneath the intruder are novel means toward the end of biting the back of the intruder. Moreover, some of a resident's means can be deployed toward novel ends. Digging is an element of the rat's behavioral repertoire that is ordinarily used for an entirely different purpose: constructing burrow systems for shelter and nesting (Boice, 1977). This suggests that there are informational states, representations of means (e.g. motor representations of digging, biting, lateral attack, etc.), that can interact interchangeably with motivational states, representations of various ends (e.g. nesting, back-biting, eating etc.), in order to produce flexible behaviors.

Isaac Wiegman
10/19/2016

Importantly, the confrontation system seems to be involved in coordinating flexible back-biting behavior. Moreover, this is something we would predict if it is a solution to the basic life problem of defending a territory from intruders. Flexibility is required to successfully repel an intruder because it is not in the intruder's best interest to be repelled easily or to act predictably. For instance, the intruder would be sure to fare poorly if it acted in a way that accommodates the attacks of the resident. So a single fixed action pattern or even a whole suite of fixed action patterns on the part of the resident would not tend to be successful against the most likely strategy of the intruder. It is more adaptive to have a flexible motivational state that leads to repeated back biting across a wide range of strategies or postures that the intruder might adopt. Rather than leading only to inflexible, stereotyped responses, it appears that solutions to basic life problems sometimes require some degree of motivational integration.

5. Implications for Emotion Theory

If we understand BEs in this way, this changes the shape of an ongoing debate in emotion theory concerning the existence of BEs in humans. In the past, this debate has carried on under the assumption that if an emotion is biologically basic, then one should predict that the various response components of the emotion will have a high degree of coherence; that for example "all instances of anger should have a characteristic facial display, cardiovascular pattern, and voluntary action that are coordinated in time and correlated in intensity."

Isaac Wiegman
10/19/2016

(Barrett, 2006, p. 29) This high degree of coherence is not observed across many emotions (Gentsch, Grandjean, & Scherer, 2013; Reisenzein, Studtmann, & Horstmann, 2013). For instance, when anger is elicited in experimental settings, it is uncommon to observe facial expressions in conjunction with the other putative components of BE anger.

One way of defending the basicality of an emotion against this criticism is to reassess what patterns of emotional response are predicted by BE theory. As we saw in the section above the motivational component of a basic emotion can select novel, instrumental behaviors. Moreover, the motivational component can be indispensable for solving a basic life problem. I think we can add to this the possibility that other response components are not as indispensable as the motivational state. To see this, suppose that anger in humans is a solution to basic life problems of deterring conspecifics from challenges and insults. If so, it may be that the only reliable requirement of successful deterrence (at least in our lineage) is a flexible motivation to retaliate against perceived wrongs (e.g. McCullough, Kurzban, & Tabak, 2012). For instance, a reliable disposition to garner a reputation for revenge (e.g. by avenging personal offenses) appears to be a highly reliable strategy for deterrence (e.g. Daly & Wilson, 1988; Frank, 1988), perhaps more so than any facial expression or physiological responses. If revenge can be served cold, then anger may not always require anything more than a motivation to avenge. If so, then we might *expect* that the only reliably occurring component of anger is the relevant motivational state. But if this is correct, then evidence of low coherence is not evidence against the existence of BE anger. While this is a just-so story that may or may not end up being true, it shows that the expected level of coherence in a BE

depends on which basic life problem shaped that emotion. In some cases, we might expect the motivational state to be the only component that does not significantly vary across the situations in which these problems arise. In that case, contextually variable responses will be the norm rather than the exception.

6. Conclusion: What Basic Emotions Really Are

So what are basic emotions? Like other theoretical terms, part of the theoretical function of basic emotions is to place selective stress on competing theories (e.g. Kroon, 1985). In this case, BEs and competing conceptions of emotion allow us to discriminate between evolutionary theories of emotion in competition with radical social constructivist theories (e.g. Barrett, 2014; Lindquist, Siegel, Quigley, & Barrett, 2013).

BEs help distinguish these theories by specifying an architecture for emotion production predicted by evolutionary considerations. The distinguishing factor is whether emotion production is categorical or dimensional (see figure 2). If each BE is a solution to a different basic life problem, then when a BE is elicited, we should see emotional responses that are relevant to that basic life problem and distinct from the responses manifested by other BEs. Emotion production is categorical in the sense that the behavioral responses are controlled by a single emotional state (as distinct from other emotional states that might control a distinct pattern of response). By contrast, if all emotions are socially constructed as

Isaac Wiegman
10/19/2016

some theorists claim, we might expect to see emotional behaviors controlled directly by multiple dimensions of appraisal (as in the bottom half of figure 2).

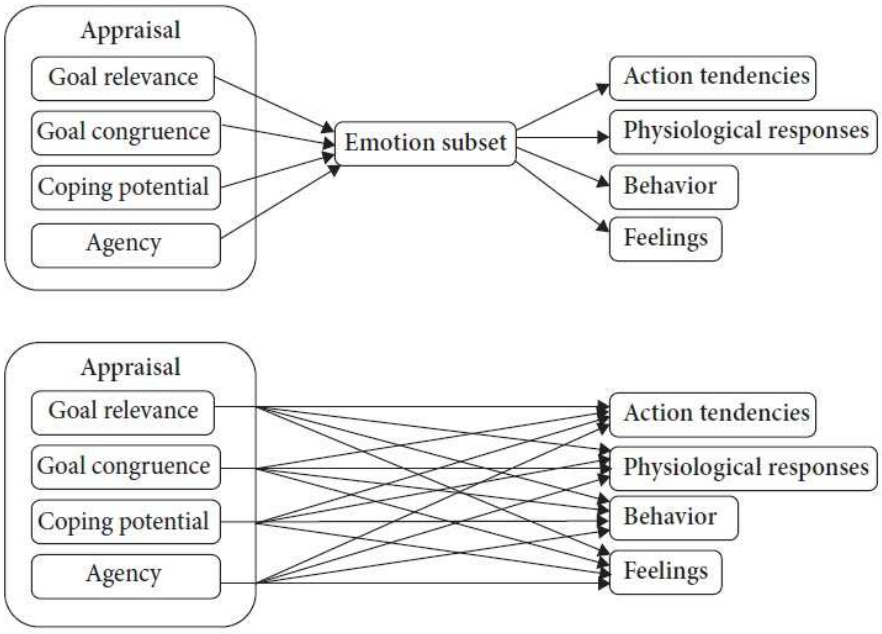


Figure 2 Competing architectures for emotion production. Top diagram is a categorical architecture, whereas the bottom is dimensional. From Moors (2012), p. 266 permission pending © John Benjamins Publishing Company. Originally published in Zachar and Ellis (2012).

Isaac Wiegman
10/19/2016

Until the present, contextual variability of emotional responses has played a decisive role in distinguishing between these two architectures for emotion production. If flexible motivational states are not included among the components of BEs, then discrete emotion production predicts insensitivity to context subsequent to elicitation (though emotion regulation processes can perhaps inhibit or augment emotional responses according to context). However, once flexible motivational states are possible, categorical emotion production is compatible with a greater amount of contextual variability.

Admittedly, this added complexity makes it more difficult to test whether humans have BEs. Nevertheless, it is not impossible. For instance, in the case of anger, researchers have developed a neurological measure of approach motivation (for a review, see Carver & Harmon-jones, 2009). If this motivational state is a component of anger, we can measure whether approach motivation itself is better predicted by contextual variables subsequent to anger elicitation or rather by contextual variables prior to or during elicitation. If contextual variables prior to elicitation do not independently predict approach motivation as BE theory might lead us to expect, then we would have evidence against the existence of BE anger.

I have argued against prevailing assumptions that BEs lack cognitive integration. In the past, evidence against cognitive integration has been concerned with informational integration, and motivational integration has not been considered. Moreover, the assumed requirements for integration concern interaction with verbally reportable or consciously accessible states, and integration with other representational states is ignored. Moreover, BEs in rodents exhibit a form of motivational integration that plausibly hinges on interaction with

Isaac Wiegman
10/19/2016

a wider variety of representational states. Properly understood, BEs are more likely to refer to emotional states in humans.

Isaac Wiegman
10/19/2016

References

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*, 407–419.

Barrett, L. F. (2006). Are Emotions Natural Kinds? *Perspectives on Psychological Science*, *1*(1), 28–58. http://doi.org/10.1111/j.1745-6916.2006.00003.x

Barrett, L. F. (2014). The Conceptual Act Theory: A Précis. *Emotion Review*, 1–20. http://doi.org/10.1177/1754073914534479

Bermúdez, J. (2003). *Thinking without words*.

Blanchard, D. C., & Blanchard, R. J. (1984). Affect and aggression: An animal model applied to human behavior. In R. J. Blanchard & D. C. Blanchard (Eds.), *Advances in the Study of Aggression* (Vol. 1, pp. 1–62).

Blanchard, D. C., & Blanchard, R. J. (2003). What can animal aggression research tell us about human aggression? *Hormones and Behavior*, *44*(3), 171–177. http://doi.org/10.1016/S0018-506X(03)00133-8

Blanchard, D. C., Litvin, Y., Pentkowski, N. S., & Blanchard, R. J. (2009). Defense and Aggression. In G. G. Berntson & J. T. Cacioppo (Eds.), *Handbook of Neuroscience for the Behavioral Sciences* (pp. 958–974). Hoboken: Wiley.

Blanchard, R. J., Blanchard, D. C., Takahashi, T., & Kelley, M. J. (1977). Attack and defensive behaviour in the albino rat. *Animal Behaviour*, *25*, 622–634.

Isaac Wiegman
10/19/2016

Boice, R. (1977). Burrows of wild and albino rats: effects of domestication, outdoor raising, age, experience, and maternal state. *Journal of Comparative and Physiological Psychology*, *91*(3), 649–61.

Carver, C. S., & Harmon-jones, E. (2009). Anger Is an Approach-Related Affect : Evidence and Implications. *Psychological Bulletin*, *135*(2), 183–204. http://doi.org/10.1037/a0013965

Daly, M., & Wilson, M. (1988). *Homicide*. Transaction Publishers.

Defensor, E. B., Corley, M. J., Blanchard, R. J., & Blanchard, D. C. (2012). Facial expressions of mice in aggressive and fearful contexts. *Physiology & Behavior*, *107*(5), 680–5. http://doi.org/10.1016/j.physbeh.2012.03.024

Eibl-Eibesfeldt, I. (1961). The Fighting Behavior of Animals. *Scientific American*, *205*, 112–122.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*. University of Nebraska Press Lincoln. http://doi.org/10.1037/0022-3514.53.4.712

Ekman, P. (1977). Biological and cultural contributions to body and facial movement. In J. Blacking (Ed.), *Anthropology of the body* (pp. 34–84).

Ekman, P. (1999). Basic Emotions. In T. Dalgleish & M. Power (Eds.), *The Handbook of Cognition and Emotion* (pp. 45–60). Sussex: John Wiley & Sons.

Isaac Wiegman
10/19/2016

Ekman, P. (2003). *Emotion Revealed: Understanding Faces and Feelings*. Phoenix Press.

Ekman, P., & Cordaro, D. (2011). What is Meant by Calling Emotions Basic. *Emotion Review*, *3*(4), 364–370. http://doi.org/10.1177/1754073911410740

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, *221*(4616), 1208–1210.

Fodor, J. (1983). *The modularity of mind: An essay on faculty psychology*.

Fokkema, D. S., Koolhaas, J. M., & van der Gugten, J. (1995). Individual characteristics of behavior, blood pressure, and adrenal hormones in colony rats. *Physiology & Behavior*, *57*(5), 857–62.

Frank, R. H. (1988). *Passions Within Reason: The Strategic Role of the Emotions*. New York: Norton. http://doi.org/10.2307/2072516

Friesen, W. (1973). *Cultural differences in facial expressions in a social situation: An experimental test on the concept of display rules.* University of California, San Francisco.

Gentsch, K., Grandjean, D., & Scherer, K. R. (2013). Coherence explored between emotion components: Evidence from event-related potentials and facial electromyography. *Biological Psychology*. http://doi.org/10.1016/j.biopsycho.2013.11.007

Ginet, C. (1990). *On action.*

Goodale, M., Milner, A., Jakobson, L., & Carey, D. (1991). A neurological dissociation

Isaac Wiegman
10/19/2016

between perceiving objects and grasping them. *Nature*.

Griffiths, P. E. (1997). *What emotions really are: The problem of psychological categories* (Vol. 1997). University of Chicago Press.

Griffiths, P. E. (2004). Emotions as Natural and Normative Kinds, *71*(December), 901–911.

Hurley, S. (2003). Animal action in the space of reasons. *Mind & Language*, *18*(3), 231–256.

Hursthouse, R. (1991). Arational actions. *The Journal of Philosophy*.

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations.

Izard, C. E. (2007). Basic Emotions, Natural Kinds, Emotion Schemas, and a New Paradigm. *Perspectives on Psychological Science*, *2*(3), 260–280. http://doi.org/10.1111/j.1745-6916.2007.00044.x

Kroon, F. (1985). Theoretical terms and the causal view of reference. *Australasian Journal of Philosophy*, (February 2014), 37–41.

Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, *27*(4), 363–384.

Lindquist, K. A., Siegel, E. H., Quigley, K. S., & Barrett, L. F. (2013). The Hundred-Year Emotion War : Are Emotions Natural Kinds or Psychological Constructions ? Comment on Lench , *139*(1), 255–263. http://doi.org/10.1037/a0029038

Isaac Wiegman
10/19/2016

McCullough, M. E., Kurzban, R., & Tabak, B. a. (2012). Cognitive systems for revenge and forgiveness. *The Behavioral and Brain Sciences*, 1–15. http://doi.org/10.1017/S0140525X11002160

Moors, A. (2012). Comparison of affect program theories, appraisal theories, and psychological construction theories. *Categorical versus Dimensional Models of Affect. A Seminar on the Theories of Panksepp and Russell*, 257–278.

Olton, D. (1979). Mazes, maps, and memory. *American Psychologist*.

Reisenzein, R., Studtmann, M., & Horstmann, G. (2013). Coherence between Emotion and Facial Expression: Evidence from Laboratory Experiments. *Emotion Review*, *5*, 16–23. http://doi.org/10.1177/1754073912457228

Rey, G. (1997). Contemporary philosophy of mind: A contentiously classical approach.

Tolman, E. (1948). Cognitive maps in rats and men. *Psychological Review*.

Zachar, P., & Ellis, R. (2012). *Categorical versus dimensional models of affect: a seminar on the theories of Panksepp and Russell*.

Multiple realization and the commensurability of taxonomies

Abstract

The past two decades have witnessed a revival of interest in multiple realization and multiply realized kinds. Bechtel and Mundale's (1999) illuminating discussion of the subject must no doubt be credited with having generated much of this renewed interest. Among other virtues, their paper expresses what seems to be an important insight about multiple realization: that unless we keep a consistent grain across realized and realizing kinds, claims alleging the multiple realization of psychological kinds are vulnerable to refutation. In this paper I argue that, intuitions notwithstanding, the terms in which their recommendation has been put make it impossible to follow, while also misleadingly insinuating that meeting their desideratum virtually guarantees mind-brain identity. Instead of a matching of grains, what multiple realization really requires is a principled method for adjudicating upon differences between tokens. Shapiro's (2000) work on multiple realization can be understood as an attempt to adumbrate such a method.

Multiple realization, *neuroscience*, *autonomy of psychology*, *intertheoretic reduction*

1. Introduction

The multiple realization ("MR") hypothesis asserts, at its baldest, that the same psychological state may be realized in neurologically distinct substrates (Polger 2009). Hilary Putnam's (1967) ingenious suggestion that pain is likely to be a multiply realized kind ("MR kind") rather neatly captures the thought here—while both mammals and molluscs presumably experience pain, they're likely to instantiate it in neurological systems of a very different sort.

MR was played against a popular philosophical theory of mind in the 1960s which attempted to identify mental states with neural states. Since MR implies a many-to-one mapping from neural states to mental states, if it is in fact true that mental states are multiply realized, it follows that no clear identity relation can hold between them. As Bechtel and Mundale (1999, 176) frame the issue, "[o]ne corollary of this rejection of the identity thesis is the contention that information

about the brain is of little or no relevance to understanding psychological processes." When the MR hypothesis first came to prominence, its critics by and large accepted it as empirically correct, and merely denied its touted antireductionist implications. In recent years the debate has struck a new note, with many philosophers calling the empirical hypothesis itself into question. Bechtel and Mundale's (1999) influential paper, followed quickly at the heels by Shapiro's (2000) penetrating analysis of functions, perhaps did most to reignite the old controversy and drag MR back into the philosophical limelight. Bechtel and Mundale express what seems to be an important insight about multiple realization: that unless we keep a consistent grain across realized and realizing kinds, claims alleging the multiple realization of psychological kinds are vulnerable to refutation. In this paper I argue that, intuitions notwithstanding, the terms in which their recommendation has been put make it impossible to follow, while also misleadingly insinuating that meeting their desideratum virtually guarantees mind-brain identity. Instead of a matching of grains, what MR really requires is a principled method for adjudicating upon differences between tokens. Shapiro's (2000) work on MR can be understood as an attempt to adumbrate such a method.

2. Bechtel and Mundale's grain requirement

Bechtel and Mundale appeal to "neurobiological and cognitive neuroscience practice" in the hope of showing how claims that psychological states are multiply realized are unjustified. Intuitively, theirs is an argument from success: cognitive neuroscience's method assumes MR is false, and the success of that method is evidence that MR *is* false. They argue that it is "precisely on the basis of working assumptions about commonalities in brains across individuals and species that neurobiologists and cognitive neuroscientists have discovered clues to the information processing being performed" (1999, 177).

Bechtel and Mundale examine both the "neuroanatomical and neurophysiological practice of carving up the brain." What they believe this examination reveals is, firstly, that the principle of psychological function plays an essential role in both disciplines, and secondly, that "the cartographic project itself is frequently carried out comparatively—across species" (1999, 177), the opposite of what one would expect if MR were "a serious option." It is the very similarity (or homology) of brain structure which permits generalization across species; and similarity in the functional characterization of homologous brain regions across

species only makes sense if the claims of MR are either false or greatly exaggerated. For instance, "[e]ven with the advent of neuroimaging, permitting localization of processing areas in humans, research on brain visual areas remains fundamentally dependent on monkey research…" (1999, 195). "The clear assumption is that the neural organization in the macaque will provide a defeasible guide to the human brain" (1999, 183). Brodmann's famous brain maps were based upon comparisons of altogether 55 species and 11 orders of mammals. If MR were true, "one would not expect results based on comparative neuroanatomical and neurophysiological studies to be particularly useful in developing functional accounts of human psychological processing" (1999, 178). They also argue that the ubiquity of brain mapping as a way of decomposing cognitive function points to the implausibility of the MR thesis. The understanding of psychological function is increasingly "being fostered by appeal to the brain and its organization" (1999, 191), again, the opposite of what one would expect "[i]f the taxonomies of brain states and psychological states were as independent of each other as the [MR] argument suggests" (1999, 190-91).

In light of such considerations, Bechtel and Mundale (1999, 178-79, 201-04) resort to grains as a way of making sense of what they perceive to be the

entrenched, almost unquestioning consensus prevailing around MR. They think that it can be traced to the practice of philosophers appealing to different grain sizes in the taxonomies of psychological and brain states, "using a coarse grain in lumping together psychological states and a fine grain in splitting brain states." When Putnam went about collecting his various specimens of pain, he ignored the many likely nuances between them. At the same time, he had few compuctions about declaring them different at a neurological level. His contention that pain is likely to be an MR kind can only command our respect if we can be sure that when he was comparing his specimens from a neurological point of view he was careful to apply no less lenient a standard of differentiation than he applied when comparing his specimens from a psychological point of view. Bechtel and Mundale maintain that when "a common grain size is insisted on, as it is in scientific practice, the plausibility of multiple realizability evaporates." As their examples of neuroanatomical and neurophysiological practice attest, scientists in these fields typically match a coarse-grained conception of psychological states with an equally coarse-grained conception of brain states. Despite the habit of philosophers individuating brain states in accordance with physical and chemical criteria, a habit no doubt originating with Putnam, this is not how neuroscientists characterize them. The notion of a brain state is "a philosopher's fiction" (1999,

177) given that the notion neuroscientists actually employ is much less fine-grained, namely "activity in the same brain part or conglomerate of parts."

A not unrelated factor is that the MR hypothesis often gets presented in a "contextual vacuum." The choice of grain is always determined by context, with "different contexts for constructing taxonomies" resulting in "different grain sizes for both psychology and neuroscience." The development of evolutionary perspectives, for instance, in which the researcher necessarily adopts a coarse grain, contrasts with the much finer grain that will be appropriate when assessing differences among conspecifics:

> One can adopt either a coarse or a fine grain, but as long as one uses a comparable grain on both the brain and mind side, the mapping between them will be correspondingly systematic. For example, one can adopt a relatively coarse grain, equating psychological states over different individuals or across species. If one employs the same grain, though, one will equate activity in brain areas across species, and one-to-one mapping is preserved (though perhaps further taxonomic refinement and/or delineation may be required). Conversely, one can adopt a very fine grain,

> and differentiate psychological states between individuals, or even in the same individual over time. If one similarly adopts a fine grain in analyzing the brain, then one is likely to map the psychological differences onto brain differences, and brain differences onto psychological differences. (1999, 202)

At least among some philosophers Bechtel and Mundale's message has evidently been well received (Couch 2004; Polger 2009; Godfrey-Smith, personal communication; see also tacit approval in Aizawa and Gillett 2009, 573). Polger (2009) explains the motivation for the grain requirement in an illuminating way. Neuroplasticity has in recent times been thought to provide compelling evidence for the MR of mental states. He concludes that "contrary to philosophical consensus, the identity theory does not blatantly fly in the face of what is known about the correlations between psychological and neural processing" (2009, 470). The grains argument figures prominently in his reasoning. As he points out, it might be tempting to regard a phenomenon like cortical map plasticity—where different brain regions subserve the same function at different times in an individual's history, say, after brain injury or trauma—as an existence proof of MR. But not if the point about grains is taken to heart. It all comes down to what we mean by "*different* brain regions" subserving "the *same* function." Consider that

recovered functions are frequently suboptimal. Genuine MR would indeed require the *same* psychological state to be underwritten by different neurological states; but suboptimality is evidence of difference underlying difference, not difference underlying sameness, as MR requires:

> It's true that this kind of representational plasticity involves the "same" function being mediated by "different" cortical areas. But here one faces the challenge leveled by Bechtel and Mundale's charge that defenses of [MR] employ a mismatch in the granularity of psychological and neuroscientific kinds. If we individuate psychological processes quite coarsely—by gross function, say—then we can say that functions or psychological states are of the same kind through plastic change over time. And if we individuate neuroscientific kinds quite finely—by precise cortical location, or particular neurons—then we can say that cortical map plasticity involves different neuronal kinds. But this is clearly a mug's game. What we want to know is not whether there is some way or other of counting mental states and brain states that can be used to distinguish them—no doubt there are many. The question is whether the sciences of psychology and neuroscience give us any way of *registering the two taxonomic systems*. (2009, 467, my emphasis)

3. Problems with the grain requirement: imprecise, impracticable, and misleading

But now the question is this: what, precisely, can it mean to use a "comparable" grain, or to keep a grain size "constant," across both psychological and neurophysiological taxonomies? Polger's motivation makes a lot of sense, to be sure, but talk of "registering" taxonomies (as of *aligning* classificatory regimes, or rendering distinct scientific descriptions *commensurable*, or however else one might care to put it) doesn't shed any light on how the desideratum for consistent grains can actually be met. Since it is intended to serve in part as a methodological prescription, it's important to know what to make of this requirement—metaphors won't help us here. How, in *concrete* terms, is an investigator meant to satisfy such a condition as *this* on their research?

Perhaps it means this. Suppose you have two tokens of fruit. The science of botany (say) could deliver descriptions under which the two are classified the same (e.g. from the point of view of *species*), but also descriptions under which they come out as different (e.g. from the point of view of *varieties*). The first

description could be said to apply a coarser grain than the second. Now imagine economics coming into the picture. The science of economics can likewise deliver descriptions under which both tokens are classified the same (e.g. both are forms of tradable fresh produce) or different (e.g. one, being typically the crunchier and sweeter variety, has a lower elasticity of demand than the other). Once again, the first description could be said to apply a coarser grain than the second. Perhaps, then, we could take it that botany and economics deliver descriptions at the same grain of analysis when their judgments of sameness or difference cohere in a given case. In the example, botanical descriptions via species classification would be furnished at the same grain as economic descriptions via commodity classification, so that species descriptions in botany are "at the same grain" as commodity descriptions in economics. By the same logic, *variety* descriptions in botany would be comparable to *elasticity* descriptions in economics. Fine. But if that is all that "maintain a comparable grain" amounts to, it really does beg the question, for this is simply type-type identity by fiat. *Of course* such a recommendation will ensure that the mapping between psychology and neuroscience will be "systematic" (to use Bechtel and Mundale's term), because on this account yielding concordant judgments of similarity or difference across taxonomies is what it *means* to apply the same grain. So we haven't solved the problem: *this* version of the grain

requirement makes type-type identity a fait accompli, effectively obliterating all MR kinds from the natural order.

It's just as well that I don't think this is what Bechtel and Mundale had in mind when they made their move to grains; supposing otherwise would serve only to trivialize an important aspect of their analysis. Still the construal is by no means far-fetched: "[o]ne can adopt either a coarse or a fine grain," they tell us, "but as long as one uses a comparable grain on both the brain and mind side, the mapping between them will be correspondingly systematic" (note that—it *will* be!). This sounds like someone with the utmost confidence in the grain requirement, which is of course what one *would* have if one thought grains could be legitimately matched in just this way. My guess is that, while they do have something important to tell us about MR, a beguiling metaphor has led them to suppose that MR is easier to refute than it actually is. (I'll support this contention with a few examples in a moment.)

Of course matters aren't much helped by the reasonable suspicion that MR is the result of pairing *inconsistent* grains. For what is neuroscience if not a fine-grained description of psychology, and psychology if not a coarse-grained

description of neuroscience? It is surely plausible that the neural and psychological sciences line up in something like this way, given that talk about the mind is really talk about the brain from a somewhat more abstract point of view.

What Bechtel and Mundale are ultimately trying to convey through their discussion of grains is the thought that claims of MR cannot be advanced willy-nilly—that there is an objective and standard way to go about verifying the existence of MR kinds and arbitrating disputes involving them. For the reasons just canvassed, however, it strikes me that talk of grains doesn't serve their purposes at all well. In fact they would have been nearer the mark had they said that what MR requires is some sort of principled *mismatching* of grains.

So far I've tried to indicate in what respects Bechtel and Mundale's grain requirement is imprecise and impracticable. Before I can show that the grains strategy is also misleading, and indeed often gets things wrong, I need to set it against an account which demonstrably gets things right.[1] Shapiro (2000) expresses with enviable lucidity what I think is the crucial insight towards which Bechtel

1 It is an account which even its detractors concede gets at least the essential point of interest to us here right, e.g. Gillett (2003).

and Mundale were uneasily groping. Interestingly, some philosophers—e.g. Polger (2009)—write as if the grain requirement and Shapiro's own formula for MR were effectively interchangeable. This is a mistake: the two approaches deliver different judgments in nontrivial cases (as I'll illustrate in a moment).

As Shapiro reminds us:

> Before it is possible to evaluate the force of [the MR thesis] in arguments against reductionism, we must be in a position to say with assurance what the satisfaction conditions for [the MR thesis] actually are. (2000, 636)

For him, "[t]he general lesson is this. Showing that a kind is multiply realizable, or that two realizations of a kind are in fact distinct, requires some work" (2000, 645). Furthermore, "[t]o establish [the MR thesis], one must show that the differences among purported realizations are causally relevant differences" (2000, 646). Shapiro's concerns revolve around what motivates ascriptions of difference, and therefore sameness. The issue is important because the classic intuition pump that asks us to conceive a mind in which every neuron has been replaced by a silicon chip depends on our ascription of an interesting difference between neurons and

silicon chips, apparently even where silicon chips can be made that contribute to psychological capacity by one and the same process of electrical transmission. His answer too, like Bechtel and Mundale's, depends ultimately on context—in particular, the context set by the very inquiry into MR itself.

Shapiro (2000, 643-44) argues that "the things for which [the MR thesis] has a chance of being true" are all "defined by reference to their purpose or capacity or contribution to some end." This is the reason why carburetors, mousetraps, computers and minds are standard fare in the literature of MR. They are defined "in virtue of what they do," unlike, say, water, which is typically defined by what it is, i.e. its constitution or molecular structure, and accordingly *not* an MR kind. Genuine MR requires that there be "*different* ways to bring about the function that defines the kind." Truly distinct (indeed *multiple*) realizations are those that "differ in causally relevant properties—in properties that make a difference to how [the realizations] contribute to the capacity under investigation." Two corkscrews differing only in color are not distinct realizations of a corkscrew, because color "makes no difference to their performance as a corkscrew." Similarly, the difference between steel and aluminium is not enough to make two corkscrews that are alike in all other respects two different realizations of a corkscrew "because, relative to

the properties that make them suitable for removing corks, they are identical." In this instance, differences of composition can be "screened off." Naturally there may be cases where differences of composition *will* be causally relevant (and it turns out that this will be important to the broader point I make below about where the grains strategy goes wrong). Perhaps rigidity is the allegedly MR kind in question. In that event, compositional differences will necessarily speak to how aluminium and steel achieve this disposition. The crucial thing to note here is that MR *is* the context, and MR makes *function* the relevant consideration, i.e. the specific point of view from which we will compare a set of tokens in the first instance (not phenomenology, not behavioral ecology, or anything else for that matter). Explanatory considerations may of course fine-tune the *sort* of function that captures our attention (cork-removal, rigidity, vision, camera vision, etc.). But function here is our key preoccupation, and having settled on a specific function which a set of tokens can be said to perform, the all-important question on Shapiro's analysis is *how* the two tokens bring that function about. Each case must be judged on its own merits. Thus unlike the two corkscrews identical in all respects save color, which do not count as distinct realizations, waiter's corkscrews and winged corkscrews are enabled to perform the same task in virtue of *different* causally relevant properties, and therefore *do* count as genuinely distinct

realizations of a corkscrew, one based on the principle of simple leverage, the other relying on a rack and pinions (Fig. 1).

(a) **(b)**

Figure 1. A waiter's corkscrew (a) and a winged corkscrew (b). Each contributes to the capacity of cork-removal in different ways.

Notice that to the extent Shapiro's causal relevance criterion envisages certain realizing properties being "screened off" from consideration in the course of inquiry, there is a sense in which the taxonomies of realized and realizing kinds may be said to be "commensurable" or "registrable" (no doubt explaining why some philosophers have simply confused commensurability with causal relevance). Thus when comparing the cork-removing properties of two waiter's corkscrews, compositional differences will not feature in the realizing taxonomy (if we accept Shapiro's characterization of the problem). So we have *cork-removal*,

which features in what we may regard as a coarse-grained taxonomy, realized by two objects described by a "science" of cork-removal in which microstructural variations do not matter, hence which might also be regarded as a coarse-grained taxonomy. If on the other hand we were comparing the same corkscrews for rigidity, where one was made of steel and the other of aluminium, compositional differences *would* feature in the realizing taxonomy. Here we would have *rigidity*, which features in what we could well regard as a more fine-grained taxonomy than that encompassing cork-removal, realized by two objects described by a science in which microstructural variations really *do* matter (namely metallurgy), and which might also be regarded as a fine-grained taxonomy, at least more fine-grained than the fictitious science of cork-removal. But my point is this: commensurability nowhere appears as an independent criterion of validity in Shapiro's account of MR, for it is an artifact of the causal relevance criterion, not a self-standing principle. Taxonomic commensurability is in fact an *implicit* requirement of the causal relevance criterion in the sense that it's taken care of once the proper question is posed. As an explicit constraint it is a will-o'-the-wisp.

Armed with this analysis, let's examine how Bechtel and Mundale attempt to refute the status of hunger as an MR kind. Putnam (1967) had compared hunger

across species as diverse as humans and octopuses to illustrate the likelihood that some psychological predicates are multiply realizable. On the basis of their grains critique, however, Bechtel and Mundale suggest that hunger will not do the work Putnam had cut out for it; for "at anything less than a very abstract level," hunger is different in octopuses and humans (1999, 202). The thought is that a finer individuation of hunger refutes the existence of a *single* psychological kind, hunger, which can be said to cross-classify humans and octopuses. Thus they essay to challenge the cognitive uniformity which MR requires at the level of psychology.

Perhaps we might first note that when identifying a *single* psychological state to establish the necessary conditions for MR, nothing Bechtel and Mundale say actually *precludes* the choice to go abstract. If context is what fixes the choice of grain (as they are surely right to point out), who's to say that context couldn't fix the sort of grain that makes hunger relevant in an abstract sense? It may be tempting to think that a more detailed description of something is somehow more *real*. But there is of course nothing intrinsically more or less real about a chosen schema relative to others that might have been chosen. There is no reason to suspect, for instance, that a determinate has any more reality than a determinable.

And yet there is a deeper problem with Bechtel and Mundale's deployment of the grains strategy here. To repeat their complaint: "at anything less than a very abstract level," hunger is different in octopuses and humans. But now why should *this* be relevant? Who would deny it? They themselves seem to be oblivious to the context which the very inquiry into MR makes paramount. They are not right to allege, as they do, that "the assertion that what we broadly call 'hunger' is the same psychological state when instanced in humans and octopi has apparently been widely and easily accepted without specifying the context for judging sameness" (1999, 203). The reason why hunger, pain, vision and so on were all taken for granted—assumed to be uniform at the cognitive level—is because MR made *function* the point of view from which tokens were to be compared. As Shapiro reminds us, "the things for which [the MR thesis] has a chance of being true" are all "defined by reference to their purpose or capacity or contribution to some end." It was understood that, say in the case of pain, regardless of phenomenal, ecological or behavioral differences between human and octopus pain (I doubt any of which were lost on Putnam), all instances of pain in these creatures had something like *detection and avoidance* in common. This might be to cast pain at "a very abstract level," but this just happens to be the context which

the inquiry into MR itself sets. A similarly abstract feature is what unites all instances of hunger: let's call it *nutrition-induction*. It is not that decades of philosophers had simply forgotten to specify the point of view from which these psychological predicates were being considered: it is rather that they simply didn't need to, since all of them had read enough of Putnam and the early functionalists to know what they were about. Phenomenal and other differences that one might care to enumerate between these predicates come a dime a dozen. But the whole point of functionalism was to abjure the inquiry into essences and focus instead on the causal role of a mental state within the life of an organism. Yes, this is to compare tokens from an "abstract level," but that's what made functionalism intriguing to begin with. And if Shapiro's analysis is any guide, it is really the *next* step in the endeavor to verify the existence of an MR kind that is the crucial one. Genuine MR requires that there be "*different* ways to bring about the function that defines the kind." So the follow-up question concerns *how* the relevant organisms achieve their detection and avoidance function, or nutrition-induction function, or whatever the case may be. It is in fact only by asking this next question that we can appreciate just how badly the grains strategy fares. The attempt to individuate hunger more finely does *not* refute the multiple realizability of hunger as between humans and octopuses. For, relative to the shared function of nutrition-induction,

it is extremely likely that humans and octopuses realize this capacity in different ways. The attempt to individuate pain more finely would likewise *not* refute the multiple realizability of pain as between humans and octopuses. For, relative to the shared function of detection and avoidance, it is extremely likely that humans and octopuses realize this capacity in different ways. So we see that the grains strategy, to the extent that it involves fine-graining psychological states in order to undermine the cognitive uniformity required by MR, sets itself a very easy job indeed, and mischaracterizes the nature of MR by its neglect of function. Moreover Shapiro's causal relevance criterion—which honors the core concerns motivating Bechtel and Mundale's resort to grains—does *not* demonstrate that hunger (or pain) is type-reducible.

A good illustration of the grains strategy in action is provided by Couch's (2004) attempt to refute the claim that the human eye and the octopus eye are distinct realizations of the kind *eye*. Conceding differences at a neurobiological level, the strategy again involves challenging the alleged uniformity at the cognitive level. As he explains, "[e]stablishing [MR] requires showing that…the physical state types in question are distinct [and] that the relevant functional properties are type identical. Claims about [MR] can be challenged at either step"

(2004, 202). Reminding us that psychological states "are often only superficially similar," and that "at a detailed level the neural differences make for functional differences" (2004, 203), he states:

> Psychologists sometimes talk about humans and species like octopi sharing the same psychological states. However, they also recognize that there are important differences involved depending on how finely one identifies the relevant features...Establishing multiple realization requires showing that the same psychological state has diverse realizations. But we can always disagree with the functional taxonomy, and claim there are psychological differences at another level of description. (2004, 203)

Thus he relates that while the two types of eyes have similar structure in certain respects, both consisting of a spherical shell, lens and retina, they use different kinds of visual pigments in their photoreceptors, as well as having different numbers of them, the octopus having one in contrast to the human eye which has four. They also have different retinas. The human retina, with rods and cones, focuses light by bending the lens and so changing its shape. The octopus eye, with rhabdomeres instead of rods and cones, focuses light by moving the lens

backwards and forwards within the shell. All these factors show up as differences in output, not just structure. The octopus, having only a single pigment, is colorblind, while its receptor's unique structure allows it to perceive the plane of polarized light. Retinal differences likewise make for functional differences, with very little information processing occurring on the octopus's retina, unlike the case of the human retina. This produces differences in stimuli and reaction times. So the two eyes might be similar, but when described with a suitably fine grain, he contends, they come out type distinct. In the result they are both physically *and* cognitively diverse, and so not genuine examples of MR.

Notice again that, contrary to what is claimed, it has not been demonstrated that type-type identity prevails here after at all (on the understanding that the kind camera $\text{eye}_{\text{human}}$ reduces to *its* distinct neural type, and the kind camera $\text{eye}_{\text{mollusc}}$ in turn reduces to *its* distinct neural type). If anything what this foray into mollusc visual physiology succeeds in showing is that, relative to the kind camera eye, human camera eyes and octopus camera eyes count as distinct realizations(!), for, assuming Shapiro's causal relevance criterion applies, human camera eyes achieve the function of *camera vision* differently to the way octopus camera eyes

achieve this function. Were we to attend to the original inquiry, which concerned whether human eyes and octopus eyes count as distinct realizations of the kind eye, Shapiro's own response, for what it's worth, is clear (2000, 645-46): here we do seem to confront a genuine case of type-type identity, as Putnam himself assumed, because, relative to the function of *vision* (not *camera vision*), both humans and molluscs achieve the function the same way (namely, by camera vision!). Differences that would be relevant at the neural level between humans and molluscs when asking how camera vision is achieved can be conveniently screened off when the question is how vision, as distinct from camera vision, is achieved. Again if pain or hunger were the kind in question, it seems more likely than not that we *would* confront a case of MR (unlike with vision), as we conjectured earlier. Explanatory context dictates the function of interest, and the function is one that we have to assume is common to the tokens in question in order to get the inquiry into MR off the ground. Indeed if Shapiro's analysis is correct, with MR we're always asking how some common function is achieved by different tokens that *do that thing*. Where there is no common function the question of MR cannot so much as arise. The fact that the question *does* arise in all the cases we've considered is a powerful indication that we're dealing with functions which all the relevant tokens actually share. The grains strategy confuses matters by suggesting that in many

cases involving putative MR kinds, psychological states can be individuated using a finer grain of description. But if what I have been saying is right, this is not the proper way to refute a putative case of MR.

That mine is the correct assessment of the situation is not only attested to by Shapiro's analysis of MR, but also by the fact that it avoids the very mug's game Polger sought to eschew by embracing the grains strategy in the first place. If for any putative MR kind I am free to cavil with the choice of your size of grain ("oh, that's far too coarse for psychology," or "now that's really not coarse enough for neuroscience"), how is the resulting game any less of a mug's game than the one we were trapped in at the start? I myself have played a few of these games with philosophers. No one wins. Couch's remarks are telling: "we can always disagree with the functional taxonomy, and claim there are psychological differences at another level of description." So the game goes on.

4. Conclusion

In sum, I think there's a genuine problem with the grain requirement. The central difficulty is that in the terms in which it's been put it is largely unworkable, and at

best no more than a loose metaphor. For a recommendation intended to serve at least in part as a methodological reform, this is clearly unsatisfactory. I don't deny that Bechtel and Mundale were onto something. But whatever value their insight into MR might have has been obscured by their unfortunate formulation of the issue. Moreover, as I have tried to show, the formulation is unfortunate not *just* because it happens to be unworkable. More worryingly, the argument from grains distorts the truth about MR by encouraging the view that mind-brain identity comes for free once we invoke the "same grain" of description across both realized and realizing kinds. But when the insight to which this locution seems to point is expressed in terms that are intelligible and empirically tractable (namely, Shapiro's causal relevance criterion), mind-brain identity seems anything but a fait accompli. Grains talk makes it tempting to think MR is easier to refute than it in fact is. It is certainly true, as Bechtel and Mundale acknowledge, that context fixes the choice of grain (where by "grain" we mean the respect under which we seek to compare a set of tokens); but we are not ipso facto obliged to employ a consistent grain across realized and realizing kinds (since this is just about meaningless as far as a researcher into these matters would be concerned and raises a host of difficulties beside). Rather than matching grains, what MR really behooves us to do is to apply a principled method for adjudicating upon differences between tokens of a

functional kind. Shapiro's work on MR shows us how to approach this important task.

References

Aizawa, Kenneth, and Carl Gillett. 2009. "Levels, Individual Variation, and Massive Multiple Realization in Neurobiology." In *The Oxford Handbook of Philosophy and Neuroscience*, ed. John Bickle, 539-81. New York: Oxford University Press.

Bechtel, William, and Jennifer Mundale. 1999. "Multiple Realizability Revisited: Linking Cognitive and Neural States." *Philosophy of Science* 66(2): 175-207.

Couch, Mark B. 2004. "A Defense of Bechtel and Mundale." *Philosophy of Science* 71(2): 198-204.

Gillett, Carl. 2003. "The metaphysics of realization, multiple realizability, and the special sciences." *Journal of Philosophy* 100(11): 591-603.

Polger, Thomas W. 2009. Evaluating the evidence for multiple realization. *Synthese* 167(3): 457-472.

Putnam, Hilary. 1967. Psychological predicates. In: *Art, mind, and religion*, eds. W. Capitan & D. Merrill, 37-48. Pittsburgh: University of Pittsburgh Press.

Shapiro, Lawrence A. 2000. "Multiple Realizations." *Journal of Philosophy* 97(12): 635-54.

Interventionist Causation in Thermodynamics

Karen R. Zwier

March 2016 (Preprint)

Abstract

The interventionist account of causation has been largely dismissed as a serious candidate for application in physics. This dismissal is related to the problematic assumption that physical causation is entirely a matter of dynamical evolution. In this paper, I offer a fresh look at the interventionist account of causation and its applicability to thermodynamics. I argue that the interventionist account of causation is the account of causation which most appropriately characterizes the theoretical structure and phenomenal behavior of thermodynamics.

1 Introduction

The interventionist account of causation has been largely dismissed as a serious candidate for application in physics. For example, a dismissal of this sort is evident in the words of theoretical physicist Peter Havas:

> We are all familiar with the everyday usage of the words "cause" and "effect"; it frequently implies the interference by an outside agent (whether human or not), the "cause", with a system, which then experiences the "effect" of this interference. When we talk of the principle of causality in physics, however, we usually do not think of specific cause-effect relations or of deliberate intervention in a system, but in terms of theories which allow (at least in principle) the calculation of the future state of the system under consideration from data specified at a time t_0 (Havas 1974, 24).

And worries about the relevance of the interventionist account of causation in physics come not only from physicists, but also from philosophers—even those who favor interventionism:

> There are important differences between, on the one hand, the [interventionist] way in which causal notions figure in common sense and the special sciences and the empirical assumptions that underlie their application and, on the other hand, the ways in which these notions figure in physics (Woodward 2007, 67).

The reasons for dismissals and worries like those above are related to a common (but problematic) assumption that causation in physics has something to do with the dynamical evolution of a closed system. The problem is that, in our preoccupation with dynamical evolution and closed systems, we tend to forget and/or neglect those areas of physics for which we do *not* have complete equations of motion or for which it *doesn't make sense* to consider entirely closed systems. And it is in those areas that the dynamical view of physical causation makes less sense and interventionism finds its home.

In this paper, I propose to take a fresh look at the interventionist account of causation and its applicability to one of those neglected areas of physics: thermodynamics. I will argue that an interventionist analysis of thermodynamics succeeds where the dynamical view of physical causation fails. As I will show, all theorizing in thermodynamics requires careful definition of the "system" under consideration, which necessarily involves attending to the boundaries that enclose the system and the conditions imposed on those boundaries. Once boundaries are adequately specified, we end up with a strong distinction between the *internal* properties and processes of the system and those *external* influences that constrain the internal dynamics. It is in the distinction between internal properties and external influences that the natural fit between the structure of thermodynamic theorizing and the interventionist account of causation becomes apparent.

The plan of this paper is as follows. In section 2, I show that interventionist reasoning is inseparable from the structural foundation of thermodynamic theory. In section 3, I show how "driving forces" and their conjugate fluxes provide a rich basis for meaningful interventionist causal claims in thermodynamics. In section 4, I use the success of interventionist causal analysis in thermodynamics to make some broader concluding remarks.

2 The centrality of manipulated equilibrium

Thermodynamic theorizing is structured around the characterization of equilibrium states and the processes by which systems move from one equilibrium state to another. But just what is a thermodynamic equilibrium state?

A thermodynamic equilibrium state is the state of a system that is *not* undergoing a change (thermal, mechanical, or chemical). However, an equilibrium state is not a spontaneous occurrence. Natural thermodynamic systems are in constant flux. They engage in all sorts of interactions: they transfer heat, push and pull on one another, change their volume, and chemically react. The very idea of a thermodynamic "system", which can only be defined by the location and/or nature of its boundaries, is in itself a theoretical concept that we impose on the world in order to do thermodynamic "bookkeeping" (Dill and Bromberg 2011, 93). In order for a thermodynamic system to achieve an equilibrium state, the system must have been allowed to relax for a sufficient amount of time without the disturbing external influences of uncontrolled contact with other systems. And such a condition requires boundaries that isolate it—or

otherwise control exchanges—from other systems. Often those boundaries are put in place artificially, by human intervention.

Consider, for example, the air in an ordinary room. If we define our thermodynamic system in relation to the walls and doors of the room, we can say that the system has a fixed volume. If no massive weather change is currently occurring, we can assume that the air pressure in the room is approximately constant (not by isolation, but by contact with an external system whose pressure is approximately constant). If some kind of air conditioning system is in place and has been running for some time, we can also say that the temperature of the room is approximately constant. We can say that most of the chemical reactions occurring in the room are in a steady state and that the concentrations of various gases are relatively uniform (except perhaps for some minor concentration gradients near any plants and/or people located in the room), with equal flow into and out of the room for each type of gas. Notice, now, that even this *almost*-equilibrium state requires artificial maintenance (the rigidity of walls, contact with an exterior reservoir supplying constant pressure, the continuous work of the air conditioner, *etc.*). Stricter equilibrium states require much more careful isolation and maintenance, and true equilibrium states (which only exist in theory) require idealized boundaries (*e.g.*, perfect thermal insulators, frictionless pistons, perfectly rigid containers, *etc.*).

There is something of a tension, however, in the way that we think about equilibrium states. On the one hand, equilibrium states are the product of external conditions imposed on a system. On the other hand, once we consider those external conditions as given, a system will *naturally* or *spontaneously* tend toward the equilibrium state allowed by the constraints. But that spontaneous or natural behavior cannot be conceived of without external constraints being placed on the system in question. To even conceive of an equilibrium state, we must ask about the conditions imposed on its boundaries. What kind of walls enclose it? Permeable, semi-permeable, impermeable? Rigid or flexible? Adiabatic or conducting? There is no such thing as an equilibrium state unless the boundaries of the system are well-defined.[1] And the conditions imposed on those boundaries constitute external interventions on the system; they effectively *set* various thermodynamic variables to take on certain values. For example, conducting walls that put a system in contact with a thermal reservoir are effectively a way of *intervening* on temperature. Likewise, a semi-permeable boundary is a way of selectively *intervening* on particle concentrations in the system. (I will return to the question of how to conceive of boundary conditions as interventions on thermodynamic variables below in section 3.)

Thus, thermodynamic equilibrium states are inherently manipulated states—manipulated to be so either by human design or by some other mechanism that effectively imposes equilibrium conditions by external intervention. And these external manipulations or interventions, which impose values on certain thermodynamic variables, are entirely consistent with the concept of an intervention

[1] In fact, a system with no defined boundaries or external constraints is effectively a universe, and its fate is something like the "heat death" discussed by Thomson, Helmholtz, and Rankine.

that has been developed by Woodward (2003) and others. According to the interventionist account of causation, an intervention directly forces a variable to take on (or remain fixed at) a certain value. Furthermore, Woodward's definition of an intervention makes no reference to human action, and thus any entity or structure playing the role of setting certain variable values or holding them fixed can fulfill the requirements for intervention. For example, a cell membrane is a structure that effectively *intervenes* to maintain a certain equilibrium internal to the cell, by keeping interior and exterior pressures equal and by maintaining certain chemical concentrations by only allowing for select passage into and out of the cell.

Now how do these manipulated equilibrium states figure into theorizing about thermodynamic processes? We begin by representing our system of interest by reference to a *thermodynamic configuration space*. The thermodynamic configuration space is the set of all possible equilibrium states of a system, where the coordinates of that space are a relatively small number of macroscopic thermodynamic variables and each point in the configuration space represents a distinct equilibrium state. For example, we might choose as coordinates the following parameters: internal energy (U), volume (V), and the particle numbers of the various species present (N_1, N_2, ..., N_i). Then the entropy function for our system, $S = S(U, V, N_1, \ldots, N_i)$, will define a hyper-surface within the configuration space (see figure 1).

With this thermodynamic configuration space and the hyper-surface defined by the entropy function in place, we can begin to theorize about any ordered sequence of states (call these $A, B, C, \ldots$) located on the hyper-surface. Notice that a curve drawn through this sequence of states looks something like a process (in fact, we call it a *quasi-static process*) in that it represents a series of changes undergone by the system. However, such a curve can be nothing like a real process, because real processes involve nonequilibrium states and the curve represents a system that remains in equilibrium along its entire length. Furthermore, the curve could never represent the *autonomous* trajectory of a system, since every state that makes up the path is an equilibrium state and no isolated system would move from one equilibrium state to another spontaneously. So in order to think about a quasi-static process as something like a process, we must think of a system being "led"—by a series of external interventions—through the succession of desired states via "hops". We effectively imagine the system being "corralled" through the sequence of equilibrium states. And by imagining the sequence of hops between states to be very small and carried out by very tiny interventions, we can approximate a smooth curve more and more closely (in fact, arbitrarily closely).[2]

In summary, the structural foundation of thermodynamic theory is the set of equilibrium states and the quasi-static "processes" that can be drawn like lines through the space of such states. As I have argued here, the very idea of an equilibrium state is not possible without reference to boundaries and the constraints that *set* the value of certain thermodynamic variables within those

[2]My discussion here closely follows that of Callen (1985, Ch. 4).

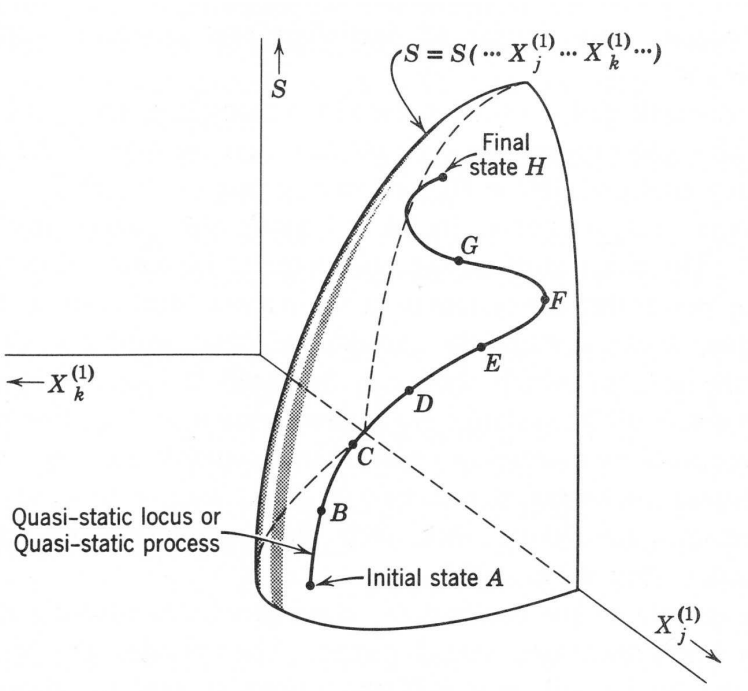


Figure 1: A representation of a quasi-static process in thermodynamic configuration space. From Callen (1985).

boundaries. Furthermore, we cannot think about quasi-static "processes", which are sequences of those equilibrium states, without thinking about a series of infinitesimal external interventions that force a system from one equilibrium state to the next. It is in this sense that interventionist reasoning is inseparable from the structural foundation of thermodynamic theory.

In the next section, I will discuss thermodynamic theorizing in greater specificity. As I will show, the interventionist view of causation maps naturally onto the use of potential functions when theorizing about a system undergoing a process.

3 Thermodynamic potentials and driving forces

The equilibrium state toward which a system will tend, given the conditions imposed on its boundaries, is governed by the energy and entropy considerations provided in the First and Second Laws of thermodynamics. The First Law tells us that any change in the internal energy (U) of a system will be equal to the total amount of energy it gains through energy exchange with the external world, in the form of heat and/or in the form of work. The Second Law tells us that any isolated system (*i.e.*, any closed system with fixed internal energy)

will tend toward its state of maximum entropy (S). The Second Law also has the result that the internal energy of any closed system with fixed entropy will be minimized. However, neither internal energy nor entropy are directly measurable, nor do we have a specific function that tells us their dependence on other state variables. What we do have, however, are other equations of state (*e.g.*, the ideal gas law) in addition to equations for U and S in *differential* form, which tell us about the way in which small changes in other state variables relate to small changes in energy and entropy:

$$dU = TdS - pdV + \sum_j \mu_j dN_j \tag{1}$$

$$dS = \left(\frac{1}{T}\right) dU + \left(\frac{p}{T}\right) dV - \sum_j \left(\frac{\mu_j}{T}\right) dN_j, \tag{2}$$

where T is absolute temperature, p is pressure, V is volume, μ_j is the chemical potential for species j, and N_j is the number of particles for species j. The above equations (and other variant forms) are commonly referred to as *thermodynamic potential functions*.

Notice that each term in both equations above involves a pair of conjugate variables. The second term in equation 1, for example, involves pressure and volume as a conjugate pair. For every pair of conjugate variables, one of the variables is extensive (*i.e.*, additive such that the property of a system is equal to the sum of that property for all of its component subsystems), while the other is intensive (*i.e.*, independent of the size of the system). Looking again at the second term in equation 1 as an example, pressure is the intensive variable and volume is the extensive variable.

Depending on the factors controlled in a given experimental context, each pair of conjugate variables tells us something about a tendency of the system as it moves toward equilibrium in that context. Since conjugate variables will be extremely important for our purposes here, let's concentrate on one pair and use an example to decipher its practical meaning.

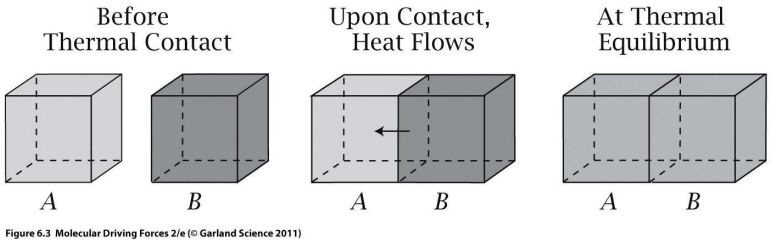


Figure 2: Two thermodynamic systems A and B before, during, and after arriving at thermal equilibrium. From Dill and Bromberg (2011, 100).

Consider the term $\left(\frac{1}{T}\right) dU$ in equation 2 and the process pictured in figure 2. We begin with two systems A and B, each enclosed in a rigid container. System A begins at temperature T_A and system B at T_B, where $T_A \neq T_B$.

The two systems are then brought into thermal contact with one another, but remain thermally insulated from the rest of the world. Now each system has an unknown entropy that can be expressed as a function of its internal energy, volume, and particle numbers, and since entropy is an extensive quantity, the total entropy of the combined system can be expressed as $S_{Total} = S_A(U_A, V_A, \mathbf{N}_A) + S_B(U_B, V_B, \mathbf{N}_B)$. Since entropy will be maximized at equilibrium, we use equation 2 to write the differential expression for S_{Total} and set it to zero:

$$\begin{aligned} dS_{Total} = \left(\frac{1}{T_A}\right) dU_A + \left(\frac{p_A}{T_A}\right) dV_A - \sum_i \left(\frac{\mu_{A_i}}{T_A}\right) dN_{A_i} + \\ \left(\frac{1}{T_B}\right) dU_B + \left(\frac{p_B}{T_B}\right) dV_B - \sum_j \left(\frac{\mu_{B_j}}{T_B}\right) dN_{B_j} = 0 \end{aligned} \tag{3}$$

If we assume that there is no particle exchange between the two systems and that no chemical change occurs within each system, we can eliminate the terms that allow for changing particle numbers. And since the containers are rigid, we can eliminate the terms that allow for changing volume. Furthermore, given that the combined system is isolated from the external world, the total internal energy of the combined system must remain constant, and any change in energy of either system must be compensated by a change in energy of the other. Thus, $dU_A = -dU_B$. So we have the following simplified expression:

$$dS_{Total} = \left(\frac{1}{T_A} - \frac{1}{T_B}\right) dU_A, \tag{4}$$

which will be equal to zero (*i.e.*, attain equilibrium) when $T_A = T_B$.

Thus we have derived the well-known result that two objects brought into thermal contact will reach equilibrium when their temperatures are equal. But more importantly for our purposes here, we can interpret the factors in equation 4 in light of this equilibration process. The difference in temperatures between the two systems leads to a nonzero value of the factor $\frac{1}{T_A} - \frac{1}{T_B}$, which effectively acts as a "force" driving a change dU_A in the internal energy of system A. More generally speaking, when a system is placed in thermal contact with a system at a different temperature, the temperature difference between the two systems acts as a force driving an exchange of heat energy between the systems. Phrased in terms of a system and its surroundings, $\frac{1}{T}$ describes the tendency of a system to exchange heat with its environment; it is the incremental relaxation that a system experiences in transferring a small bit of its energy dU.[3]

Physicists commonly use the language of "driving forces" in referring to the intensive parameters in the thermodynamic potential functions. Looking back again at equation 2, a difference between the pressure p of the system and its environment will act as a driving force for an exchange of volume dV between the system and its environment, and a difference between the concentration of a

[3] Alternatively, we could have begun with the thermodynamic potential function for internal energy (equation 1) to derive the same result.

particular species μ_j in the system and its environment will act as a driving force for exchanges of particles of the respective species with the environment (dN_j). The force or tendency represented in each of the conjugate pairs ($T, p, \boldsymbol{\mu}$) can act, separately or together (depending on the constraints imposed on the process), to drive changes in its paired extensive variable (dU, dV, or $d\mathbf{N}$, respectively), and thus to drive the system and its environment toward the equilibrium state of maximum entropy.[4]

This "driving force" language—and its basis in the way in which the environment exchanges energy and entropy with a system—matches the way in which relationships among thermodynamic variables would be modeled by the interventionist account of causation. According to the interventionist account, a variable X is an interventionist cause of another variable Y if there is a possible intervention on X that will change the value of Y (or the probability distribution over the values of Y) when the values of all other variables in the system remain fixed.[5] In physical experiments, the condition that the values of all other variables in the system remain fixed across changes in the intervention on X is often enforced using what I will call "auxiliary interventions" on those variables. To see how interventionist treatment matches the "driving force" language, let's consider the temperature equilibration case above, with system A as the causal system under investigation.

Consider the set of thermodynamic variables characterizing system A when we consider the temperature equilibration process in terms of maximization of entropy: volume V_A, the set of particle numbers for each species $\mathbf{N}_A$, temperature T_A, and internal energy U_A. Each of these variables is represented below in figure 3. The primary intervention in the temperature equilibration case was the operation of placing system B in thermal contact with system A. This intervention occurred specifically under conditions in which the volume V_A and particle numbers $\mathbf{N}_A$ of system A were held constant; the enforcement of constant values of V_A and $\mathbf{N}_A$, by enclosing the system within rigid impermeable walls, constitutes the set of auxiliary interventions in this case. Under the conditions set by these auxiliary interventions, the primary intervention produced a change in T_A, since the original temperatures of the two systems were not equal, and this change in temperature resulted in a change in the internal energy (U_A) of the system. And since, under conditions where all other variables are held constant, the intervention was an intervention on T_A and resulted in a change

[4]Physicists use the language of "driving forces" in both the entropy and energy representations. When we flip between the energy picture of a system and the entropy picture of that same system, the metric by which we measure progress toward equilibrium changes. Each metric has its own way of characterizing the driving force because, in changing our metric of progress, there is a transformation on the force term. Still, physically, it is one and the same force driving the system toward equilibrium. This representational change in the physical equations mirrors a widely-noted feature of the interventionist account of causation: when we change the set of variables with which we characterize a causal system, our characterization of the causal relationship itself can change.

[5]I have ignored some technical details for the sake of simplicity here. See Woodward (2003, 59) for the more precise interventionist criteria for X's being a type-level direct cause of Y and X's being a type-level contributing cause of Y.

in U_A, we can say that T_A is an interventionist cause of U_A.

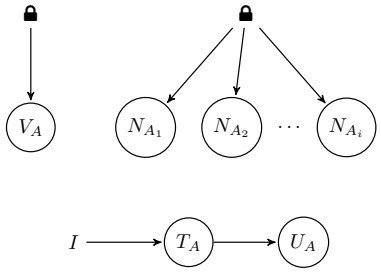


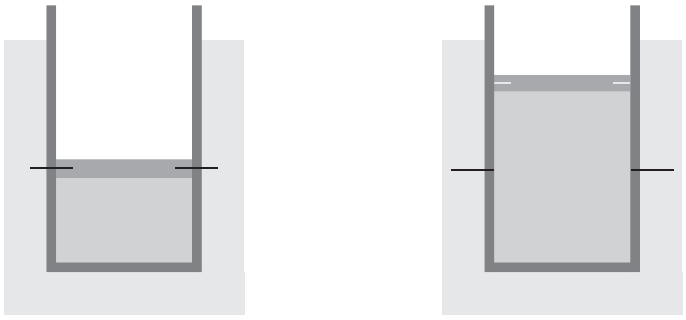
Figure 3: An interventionist causal graph of the temperature equilibration process in which system A, originally at temperature T_A, is brought into contact with another system B, originally at temperature T_B. The variable I represents the intervention that places the two systems in contact and thus changes the value of T_A. The lock symbols (🔒) represent the auxiliary interventions which hold V_A and $\mathbf{N}_A$ fixed.

To further flesh out the causal claim being represented by the arrow from T_A to U_A in figure 3, we can contrast varying interventions in which we put system A in contact with system B at varying temperatures $T_{B1}, T_{B2}, \ldots T_{Bn}$, while still holding V_A and $\mathbf{N}_A$ constant at the same values. Under such varying interventions, we will find that there are corresponding variations in the final T_A and U_A. Therefore, the interventionist account confirms that the temperature T_A of system A is a cause of its internal energy U_A. In general, interventions on temperature lead to changes in internal energy via exchange of heat when volume and particle numbers are held constant. Such a causal claim seems to be precisely what physicists mean to convey when they use "driving force" language with respect to temperature.

The intervention in the above case, where we have an equilibration process between two finite systems with differing initial temperatures, is an example of a "soft" or "parametric" intervention in that it *modifies* the temperature of our system rather than determining it completely.[6] When we put system A with its initial temperature T_A in contact with system B with its initial temperature T_B, the combined system finds an equilibrium temperature somewhere between the initial values of T_A and T_B. But thermodynamics also provides conceptual tools for theorizing about "hard" or "structural" interventions that entirely determine the value of an intensive parameter for a system. We call these theoretical entities "reservoirs" or "baths", and they have the property of being able to exchange one or more extensive quantities while their corresponding intensive properties remain constant. For example, an energy bath (*i.e.*, a temperature reservoir), by virtue of its size, is able to exchange energy with a system with which it is put in contact with negligible effect on its temperature. Likewise, a volume bath (*i.e.*, a pressure reservoir) is able to exchange volume while remaining at constant pressure, and a particle bath (*i.e.*, concentration reservoir) is able to exchange particles while maintaining constant particle con-

[6] For the distinction between soft and hard interventions, see Eberhardt and Scheines (2007).

centrations. When we theorize about cases in which we put a system in contact with a reservoir instead of a finite system, we consider a hard intervention that *determines* the value of the relevant intensive variable in our system. Such theoretical experiments bring the interventionist causal structure into even clearer relief: putting a system in contact with a reservoir is an intervention that sets the value of an intensive variable in the system, which in turn results in a change in the corresponding extensive variable.



(a) (b)

Figure 4: Illustration of a pressure-driven process, depicting (a) the system in its initial equilibrium state before the piston-locking pins are released; (b) the system once it has reached its new equilibrium state after the pins are released. This image shows the result of the case where $p_0 > p_{Res}$ and the piston rises, but all of the same considerations would apply in the case that $p_0 < p_{Res}$ and the piston falls.

Let's look at an example. Consider a system that is in an initial equilibrium state $(p_0, T, \mathbf{N})$. Suppose that we intervene on the system by bringing it into contact with a reservoir that maintains the same temperature T as the system but a different pressure p_{Res}. We might do so by releasing an initially-locked piston, allowing it to move freely between the system and the reservoir (see figure 4). The process that ensues will be ruled by a maximization of the entropy of the total combined system, so we are interested in the condition where $dS_{Total} = 0$:

$$dS_{Total} = \frac{1}{T_{Res}} dU_{Sys} + \frac{p_{Sys}}{T_{Res}} dV_{Sys} + \frac{1}{T_{Res}} dU_{Res} + \frac{p_{Res}}{T_{Res}} dV_{Res} = 0 \quad (5)$$

Due to conservation of volume and conservation of energy, $dU_{Sys} = -dU_{Res}$ and $dV_{Sys} = -dV_{Res}$, so the above condition reduces to the following:

$$dS_{Total} = \left(\frac{p_{Sys} - p_{Res}}{T_{Res}} \right) dV_{Sys} = 0 \quad (6)$$

We can see here that it is the pressure difference between system and reservoir that is driving the exchange of volume. And again, this physical interpretation in terms of driving forces matches the interventionist causal account. By placing the system in contact with the reservoir, we *set* the pressure of the system to a new value, and the forced change in pressure results in a change in volume. Were we to impose a different pressure on the system by placing it in contact

with a reservoir at a different pressure, we would see the corresponding volume change as well. Thus, pressure is an interventionist cause of volume (see figure 5).

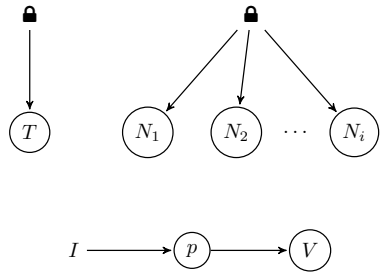


Figure 5: Interventionist causal representation of the pressure equilibration process depicted in figure 4. The variable I represents the intervention that places the system in contact with the pressure reservoir and thus changes the value of p. The lock symbols (🔒) represent the auxiliary interventions which hold T and $\mathbf{N}$ fixed.

As shown in the examples above, the most important key to successful thermodynamic theorizing is the careful definition of the boundaries between systems and accounting for the transactions that occur at those boundaries. Interventionist reasoning fits naturally into thermodynamic theorizing because its distinction between the interventions external to a causal system and the causal relations internal to that system is perfectly applicable where thermodynamic boundaries are well-defined. Since interventions are always performed *on* a causal system from outside, it is entirely natural to label exchanges between a system and its environment as interventions of the environment on those systems.

4 Conclusion

In this paper, I have shown that there is a natural fit between thermodynamic theorizing and the interventionist account of causation. I therefore argue that the interventionist account is the most suitable account of causation for describing thermodynamic theorizing and our actual interactions with thermodynamic systems.

I suggested at the beginning of this paper that we tend to assume that physical causation will have a dynamical form, and that my identification of interventionism as the most appropriate account of causation in thermodynamics would run contrary to this assumption. It might be objected that this is a somewhat dull result, however. Thermodynamics, so the objection might run, is not "fundamental" physics, and so it is unsurprising that we should find interventionist causation rather than dynamic causation in a realm of physics that is…well…*not dynamical*. But such an objection would miss the point. Our common assumption that "physical causation" must refer to the dynamical propagations of systems is the result of our preoccupation with "fundamental" physics (which

we also assume, almost by definition, must have a dynamical form) and neglect of those areas of physics which are considered to be "non-fundamental".[7]

So what is it to be a cause in (at least some of) physics? Here is a simple answer: an account of causation which appropriately characterizes the theoretical structure and phenomenal behavior of a domain of physics gives an account of what it is to be a cause in that domain of physics. And I have shown that the interventionist account does just that in thermodynamics.

References

Batterman, Robert, ed. 2013. *The Oxford Handbook of Philosophy of Physics*. Oxford: Oxford University Press.

Callen, Herbert B. 1985. *Thermodynamics and an Introduction to Thermostatics*. 2nd ed. New York: John Wiley & Sons.

Dill, Ken A. and Sarina Bromberg. 2011. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*. 2nd ed. New York: Garland Science.

Eberhardt, Frederick and Richard Scheines. 2007. "Interventions and Causal Inference." *Philosophy of Science* 74 (5): 981–995.

Havas, Peter. 1974. "Causality and Relativistic Dynamics." In *Causality and Physical Theories*, edited by William B. Rolnick, Vol. 16 of *AIP Conference Proceedings*, 23–47. New York: American Institute of Physics.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

——— 2007. "Causation with a Human Face." Chap. 4 in *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by Huw Price and Richard Corry, 66–105. Oxford: Clarendon Press.

[7] Increasingly, the study of "non-fundamental" theories has revealed that their relationship with "fundamental" theories is less straightforward than might be expected. For recent discussions in this vein, see, for example, Batterman (2013). Furthermore, it is entirely unclear what the criteria for "fundamental" status are, or whether undisputed criteria even exist. And with the criteria for fundamentality in doubt, it is hard to see what basis we might have for even expecting "fundamental" theory to always have dynamical form.

PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association

Atlanta, GA; 3-5 November 2016

Version: 29 October 2016

PhilSci ARCHIVE

Table of Contents

Page

CENTRALISED FUNDING AND THE DIVISION OF COGNITIVE LABOUR

Abstract. Project selection by funding bodies directly influences the division of cognitive labour in scientific communities. I present a novel adaptation of an existing agent-based model of scientific research, in which a central funding body selects from proposed projects located on an epistemic landscape. I simulate four different selection strategies: selection based on a god's-eye perspective of project significance, selection based on past success, selection based on past funding, and random selection. Results show the size of the landscape matters: on small landscapes historical information leads to slightly better results than random selection, but on large landscapes random selection greatly outperforms historically-informed selection.

Word count: 4359

Introduction

National funding bodies support much of contemporary science. The selection criteria for funding have gained increasing attention within philosophy of science (Gillies, 2008; O'Malley et al., 2009; Haufe, 2013; Lee, 2015). Meanwhile, there has been growing interest in model-based approaches to understanding the social epistemic activities of scientists (Kitcher, 1990; Strevens, 2003; Weisberg and Muldoon, 2009; Grim, 2009; Zollman, 2010). The current paper builds on previous modelling tools to explore the effects of centralised selection mechanisms on the division of cognitive labour and the ability of scientific communities to efficiently discover significant truths.

Science aims at discovering significant truths, i.e. not just any truths, but truths that will eventually contribute in a meaningful way to well-being (Kitcher, 2001). This is the justification for the public support of science, including basic science (Bush, 1945). Some funding terminology: scientific projects have high *impact* (ex post) if they result in significant truths; projects have high *merit* (ex ante) if they are predicted to have high impact.

Polanyi (1962) analysed merit as being composed of three components: scientific value, plausibility and originality. Polanyi notes an essential tension between plausibility and originality: the more original a project, the more difficult it is to evaluate its plausibility. Polanyi advocates selection by peer review as a conformist position, that sacrifices the occasional meritorious original project while ensuring all supported research projects are plausible, to "prevent the adulteration of science

by cranks and dabblers" (p. 8). Gillies (2008, 2014) takes an opposing position, arguing that the cost of losing (infrequent) highly original and meritorious research is much greater than the cost of occasionally supporting implausible research that ends up being of low impact. As an alternative to peer review, Gillies advocates random selection. The tension between plausibility and originality is clearly relevant to questions of effective division of cognitive labour, and has direct links to science policy. This tension, and its complexity, is explored in this paper.

I will argue that the results of the simulations presented are both significant and surprising. The simulations show that, under reasonable parameter values for at least some fields of science, choosing projects at random performs significantly better, in terms of accumulated significant truths, compared to other funding strategies, including project selection by peer review. The results support, to an extent, Gillies' proposal of funding by lottery.

1. Model description

The model explores the influence of different funding mechanisms on the accumulation of significant truths. It builds on the epistemic landscape model developed by Weisberg and Muldoon (2009), extending it by adding representations of centralised funding selection and dynamic changes in project merit. The latter is added to reflect a more realistic picture of scientific merit. For example, Strevens (2003) discusses the effect of a successful discovery on all further pursuits of the same question: they no longer have any merit, as they lose all originality. Several dynamic processes affecting merit are detailed later in the paper.

The model represents a population of scientists exploring a topic of scientific interest. They are all funded by the same central funding body to pursue projects of varying duration, measured in years. Each project's significance is allocated in advance by the modeller, from a "god's-eye" perspective. When grants end scientists successfully complete their project. Their projects' results contribute to the collection of significant truths in the field's corpus of knowledge. Funding mechanisms are compared by their ability to generate this accumulation of significant truths.

For simplicity, scientists in the model (unrealistically) do not share their findings nor explore similar projects during research. They only work on the project for which they were funded and they only share their results at the end of a grant. The social processes set aside here have been explored in previous works (Grim, 2009; Zollman, 2010). Future work may combine the different models towards a unified picture of the division of cognitive labour.

Funding is represented as a process of selection. In every time step, the scientists whose grants have run out are placed in a pool of candidates along with new entrants to the field, and the modelled funding mechanism selects from this pool of candidates those who will receive funding and carry out research projects. Modelled funding mechanisms differ in the way they select individuals, as outlined below.

Actual potential: Actual potential, which can only be known from a god's-eye perspective, is the significance of a project's results *were it successfully completed today*. In the absence of time-dependant merit, actual potential is simply the significance of the project's results. However, in the presence of time-dependence the significance could change between the initiation of the project (at the point of funding) and its completion (at the point of contributing the results to the relevant corpus). This means that in the presence of time-dependence, actual potential might diverge from the eventual contribution of the project.

Estimated potential: Estimated potential is the scientific community's ex ante evaluation (assumed, for simplicity, to be single-valued) of the merit of a proposed project. This prediction is taken to rely on the known contributions of past projects which bear some similarity to the proposed project, and so depends on the history of research projects in the field. In representing decisions based on the research community's prediction, this selection method is akin to peer-review.

Past funding: Under this mechanism, funding is allocated to those scientists who already received funding in the past, and only to them. The model (unrealistically) represents all scientists as being of equal skill, and so this mechanism cannot be taken to mean the selection of the most "intrinsically able" scientists. Rather, this mechanism is included as a "most conservative" option, not admitting any new researchers to the field beyond the field's original investigators.

Lottery: Under a lottery, all candidates have equal chances of being funded. The lottery option serves both as a natural benchmark for other funding methods, and as a representation of the mechanism proposed by Gillies (2014).

The essence of the model is the comparison of the performance of these selection mechanisms in generating results of high significance over time under various conditions.

To represent in the model the time-dependence of merit, the significance contributions of different project results are allowed to change over time as a response to scientists' actions. Three dynamic processes are included in the model (details in §2.5). Two processes involve a reduction of significance following a successful project or breakthrough,

which reflects the one-off nature of discovery. The third process involves an increase in significance when a new avenue of research is opened by a significant discovery. Simulations based on the model show that these dynamic processes have a significant effect on the relative performance of different funding strategies.

2. Simulation details

2.1. **Simulating the epistemic landscape.** To investigate the complex nature of the domain being modelled, the model was turned into a computer simulation.[1] The basic structure of the landscape simulation follows Weisberg and Muldoon's, of a two-dimensional configuration space, charted with two coordinates x and y, with an associated scalar field represented in a third dimension as height along the z axis. Each (x, y) coordinate pair specifies a different potential research project; the closer two projects are on the landscape, the more similar they are. The scalar value associated to the coordinate represents the significance of the result obtained on a successful completion of the project, were it completed today (allowing for time dependence). The limit to two spatial dimensions of variation between projects is likely to be unrealistic (Wilkins, 2008), but a higher-dimensional alternative would make the model much less tractable.

In each run of the simulation, the landscape is generated anew in the following process:

(1) Initialise a flat surface of the required dimensions.
(2) Choose a random location on the surface.
(3) Pick random values for relative height, width along x, and width along y.
(4) Add to the landscape a hill at the location chosen in step 2 by using a bivariate Gaussian distribution with the parameters picked in step 3.
(5) Repeat steps 2-4 until the specified number of hills is reached.
(6) Scale up linearly the height of the landscape according to the specified maximum height.

This process generates the "god's-eye" perspective of the research potential of the domain. Here and later, random variables are used to fill-in parameters whose existence is essential for the simulation, but where (1) the specific values they take can vary across a range of valid model targets, and/or (2) there is no compelling empirical evidence to choose a particular value. This requires, however, several runs of the simulation for each configuration, to average out the effects of random variation.

[1] Source code for the simulation is available from the author on request.

2.2. **Simulating agents.** The agents in the model represent scientists investigating the epistemic landscape. Each agent represents an independent researcher or group, and is characterised by its location on the landscape, representing the project they are currently pursuing, and a countdown counter, representing the time remaining until their current project is finished. Like Weisberg and Muldoon's "hill climbers", agents are simulated as local maximisers. Agents follow the following strategy every simulation step:

(1) Reduce countdown by 1.
(2) If countdown is not zero: remain in same location.
(3) If countdown is zero: contribute to the accumulated significance the significance of the current location, and attempt to move to the highest local neighbour.

In the simulation, the agents are identical, in the sense that any agent, when successfully completing a project of a given significance, will contribute exactly that amount to the accumulated significance of the field. This simplification ignores natural ability and gained experience, and stems from a focus on a particular approach to science funding, which funds *projects*, rather than funding *people*. The focus is informed by the explicit policies of certain funding bodies, like the National Institutes of Health (NIH), reflected, for example, in the institution of blind peer review. Thus, the results of the current work would not extend to the minority of science funding bodies, such as the Wellcome Trust, that make explicit their preference to fund people rather than projects.

The *local neighbourhood* of an agent is defined as the 3×3 square centred on their current position. The attempt to move to the highest neighbour depends on the selection (funding) mechanism, as discussed below. The *accumulated significance*, which is the sum of all individual contributions to significance, is stored as a global variable of the simulation and used to compare strategies.

In the beginning of the simulation, a specified number of agents are seeded in random locations on the landscape, with randomly generated countdowns selected from a specified range of values. An example of an initial seeding of agents can be seen in Fig. 1.

In the absence of selection and time-dependence, the course of the simulation is easy to describe: agents begin in random locations on a random landscape, and as the simulation progresses the agents finish projects and climb local hills, until, after an amount of time which depends on the size of the landscape, the number and size of peaks, and the duration of grants, all agents trace a path to their local maxima and stay there. Since agents increase their local significance during the climb, the rate of significance accumulation increases initially, until all agents reach their local maxima, at which point significance continues accumulating at a fixed rate indefinitely. This is the dynamic

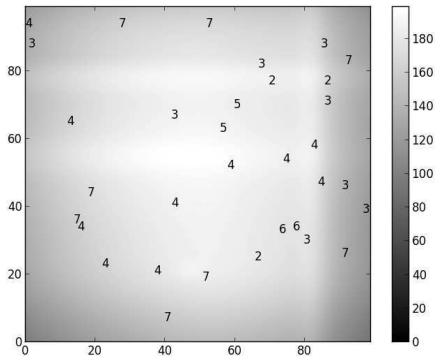


FIGURE 1. Landscape simulation with initial seeding of agents. Each number on the landscape represents an agent at its location, with the value of the number representing the agent's countdown. The colours indicate the height (significance) of each position (project) in the landscape.

seen in Weisberg and Muldoon's simulation for a pure community of "hill climbers", and its unrealistic nature highlights the importance of simulating the time-dependence of significance.

2.3. **Simulating communal knowledge.** In addition to their contribution to significance, agents also contribute to the *visibility* of the landscape (Muldoon and Weisberg, 2011). The visibility of a project represents whether the scientific community, and especially funding bodies, can estimate the significance contribution of that project. Initially, the entire landscape is invisible, representing full uncertainty. Upon initial seeding of agents, each agent contributes vision of their local neighbourhood, as defined above, to the total vision. As the agents move, they add vision of their new local neighbourhood. Visibility is used in the *best_visible* funding mechanism described below.

The simulation represents visibility in a simplistic manner by assigning binary values: either the community knows what the significance of a project will be, or it does not. A more realistic representation will allow partial visibility, with some distance decay effect, such that the community would still be able to make predictions of significance for less familiar projects, but these predictions will have a probability of being wrong, with the probability of error increasing the more unfamiliar these projects are. This addition, however, will be computationally heavy, as it requires maintaining multiple versions of the landscape, both for the real values and for the estimated values.

2.4. **Simulating funding strategies.** The aim of the model is to explore the effects of funding mechanisms on the population and distribution of investigators. Since the aim is to simulate current funding practices (albeit in a highly idealised manner), and since current funding practices operate in passive mode (choosing from proposals originating from scientists rather than dictating which projects aught be pursued), the guiding principle of the simulation is that a funding mechanism is akin to a selection process: at each step of the simulation, the actual population of agents is a subset of the candidate or potential population, where inclusion in the actual population follows a certain selection mechanism.

Funding mechanisms are simulated in the following manner:
Every step:

(1) Place all agents with zero countdown in a pool of "old candidates".
(2) Generate a set of new candidate agents, in a process identical to the seeding of agents in the beginning of the simulation.
(3) Select from the joint pool of (old candidates + new candidates) a subset according to the selection mechanism specified by the funding method.
(4) Only selected agents are placed on the landscape and take part in the remainder of the simulation, the rest are ignored.

The simulation can represent four different funding mechanisms:

best: selects the candidates which are located at the highest points, regardless of the visibility of their locations. This simulates a mechanism which selects the most promising projects from a god's eye perspective. This overly optimistic mechanism does not represent a real funding strategy. Rather, it serves as an ideal benchmark against which realistic funding mechanisms are measured.

best_visible: filters out candidates which are located at invisible locations, i.e. candidates who propose to work on projects which are too different from present or past projects. It then selects the candidates in the highest locations from the remainder. This strategy is closer to a realistic representation of selection by peer review. Note that even this version is epistemically optimistic, as it assumes the selection panel has successfully gathered all available information from all the different agents, both past and present.

lotto: selects candidates at random from the candidate pool, disregarding the visibility and height of their locations.

oldboys: represents no selection: old candidates continue, no new candidates are generated.

The key parameters for all funding mechanisms are the size of the candidate pool and the size of the selection pool. The size of the candidate pool, which in turn depends on the size of the new candidate pool (as the size of the old candidate pool emerges from the simulation), has been chosen in the simulations such that the total candidate pool is equal in size to the initial number of agents (except *oldboys* where there are no new candidates). This means the success probability changes between funding rounds, around a mean which is equal to 1/(average countdown). With an average grant duration of five years, this yields a success rate of 20%, close to the real value in many contemporary funding schemes (NIH, 2014). The number of grants awarded each year is set to equal the number of grants completed each year, maintaining a fixed size for the population of investigators.

For simplicity, the simulated funding mechanisms do not take into account the positions of existing agents on the landscape, except indirectly when considering their vision. Future simulations may consider a selection mechanism which explicitly favours either diversity or agglomeration, though one expects difficulties in operationalisation and measurement of epistemic diversity.

2.5. **Simulating merit dynamics.** To make the simulation more realistic, the significance of projects is allowed to change over time in response to research activities of the community of investigators. Three such dynamic processes are included in the simulation:

Winner takes it all: As was made explicit by Strevens (2003), the utility gain of discovery is a one-off event: the first (recognised) discovery of X may greatly contribute to the collective utility, but there is little or no contribution from further discoveries of X. In the simulation, this is represented by setting the significance of a location to zero whenever an agent at that location has finished their project and made their contribution to accumulated significance. This effect is triggered whenever any countdown reaches zero, which makes it quite common, but it has a very localised effect, only affecting the significance of a single project.

Reduced novelty: When a researcher makes a significant discovery, simulated by finishing a project with associated significance above a certain threshold, the novelty of nearby projects is reduced, which in the model is simulated by a reduction of significance in a local area around the discovery.

New avenues: When a researcher makes a significant discovery, it opens up the possibility of new avenues of research, simulated in the model by the appearance of a new randomly-shaped hill at a random location on the landscape.

3. Results and discussion

Here I present the results of simulations of different setups of interest, exploring the relative success of different funding mechanisms under different conditions.

All simulation results show a comparison between the four funding mechanisms, as a plot of total accumulated significance (arbitrary units) at the end of the simulation run, averaged over five runs with different random seeds. In all simulations the range of countdowns was 2 to 7. The number of individuals was set to equal (size of landscape)$^{3/4}$. Simulations were ran for 50 steps. The trigger for significance-dependant processes was 0.7 of the global maximum. Results are shown for a small landscape (50×50) in Fig. 2 and for a large landscape (500×500) in Fig. 3.

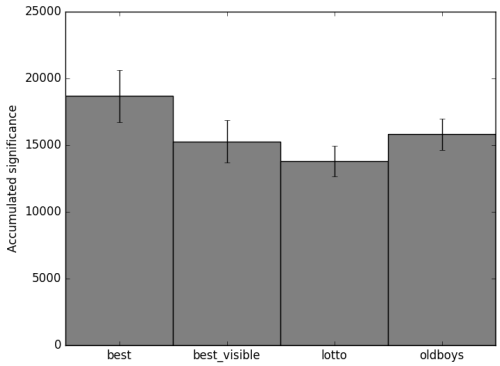


Figure 2. Comparison of significance accumulation under different funding mechanisms, small landscape (50×50).

To get a feeling for how the community is affected by the funding mechanism, I present visualisations of the state of the landscape at the end of the simulation run for the two funding mechanisms mentioned in the introduction (*best_visible* and *lotto*) in Fig. 4. Note that due to the *winner takes it all* dynamic process it is possible to "see" the past trajectory of exploration, as completed projects leave behind highly localised points of zero (remaining) significance. This allows for a visual representation of the division of cognitive labour that emerges under different funding schemes.

As is clear from the simulations, the *best* funding mechanism is indeed best at accumulating significance over time, though with various lead margins over the second best strategy. In the presence of dynamic

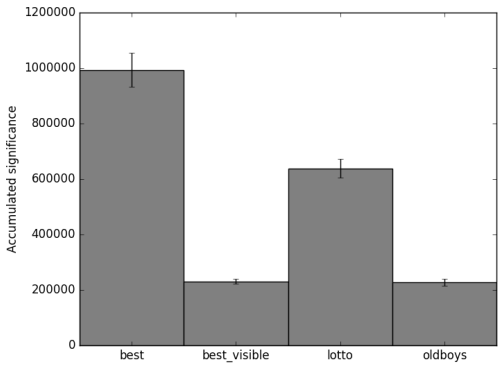
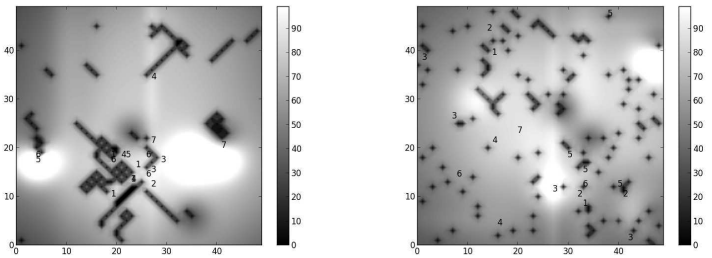


FIGURE 3. Comparison of significance accumulation under different funding mechanisms, large landscape (500×500).



(A) *best_visible* (B) *lotto*

FIGURE 4. Landscape visualisation at the end of the simulation run under different funding mechanisms.

processes, *best* is in the best position to locate new avenues for research, wherever they show up. However, as mentioned above, the *best* funding strategy is not realisable, as it requires a god's eye view of the epistemic landscape.

On the small landscape the three strategies, *best_visible*, *oldboys*, and *lotto* perform roughly similarly, with *lotto* at a small disadvantage as it cannot make use of valuable information from past successes. It seems counter-intuitive that *best_visible* performs worse than *oldboys*. A possible explanation is the effect of reduced novelty: *best_visible* tends to cluster scientists around the most promising projects, and so when one makes a breakthrough it reduces the significance of contributions for all groups working on similar projects (the phenomenon known in

contemporary science as "scooping"). This excessive clustering around fashions is not present in *oldboys* or *lotto*.

On the large landscape *lotto* greatly outperforms *best_visible* and *oldboys*. This is because new avenues on a large landscape are likely to spawn outside the visibility of the agents, where *lotto* can access them but the other two strategies cannot. In the smaller landscape this effect is not apparent, as the relative visibility is larger, and therefore the chance of a new avenue appearing within the visible area is larger.

Conclusion

This paper presented a way to extend existing epistemic landscape models so that they can represent selection by a central funding body and time dependence of significance. This model was used in computer simulations to compare the effectiveness of different idealised versions of selection criteria, most notably selection based on past successes (akin to peer review), random selection and no selection. The most significant result from the simulation was that on a large landscape, when a topic can be explored in many ways that could be very different from each other, random selection performs much better than selection based on past performance.

This result fits in with a general result from the body of works on agent-based models of scientific communities, that shows diversity in the community trumps individual pursuit of excellence as a way of making communal epistemic progress. The tension of science funding, between originality and plausibility, is thus a part of the broader tension between diversity and excellence, between exploration and exploitation.

Previous social epistemology models have focused on the role of *internal* factors in shifting the balance between exploration and exploitation. Kitcher (1990); Strevens (2003) look at reward structures (of internal credit, not external monetary rewards) and individual motivation towards credit or truth. Grim (2009); Zollman (2010) look at information availability and information transfer between scientists, and at individual beliefs. Weisberg and Muldoon (2009) look at individual researchers' social strategy: follower or maverick.

The current work is the first within this modelling lineage to look at the effects of an *external*, *institutional* factor: selection by a centralised funding body. The current paper brings this line of research closer to having a direct relevance to science policy. Hopefully future work in this vain will continue this trend, to deliver on the challenge set out by Kitcher (1990, p. 22):

> How do we best design social institutions for the advancement of learning? The philosophers have ignored the social structure of science. The point, however, is to change it.

We could start by advocating for funding mechanisms that allow for more exploration.

References

Bush, V. (1945). *Science, the endless frontier: A report to the President*. Washington: U.S. Government printing office.

Gillies, D. (2008). *How should research be organised?* London: College Publications.

Gillies, D. (2014). Selecting applications for funding: why random choice is better than peer review. *RT. A Journal on research policy and evaluation 2*(1).

Grim, P. (2009). Threshold phenomena in epistemic networks. In Complex adaptive systems and the threshold effect: Views from the natural and social sciences: Papers from the AAAI Fall Symposium, pp. 53–60.

Haufe, C. (2013). Why do funding agencies favor hypothesis testing? *Studies in History and Philosophy of Science Part A 44*(3), 363–374.

Kitcher, P. (1990). The division of cognitive labor. *The journal of philosophy 87*(1), pp. 5–22.

Kitcher, P. (2001). *Science, truth, and democracy*. New York: Oxford University Press.

Lee, C. J. (2015). Commensuration bias in peer review. *Philosophy of Science 82*(5), 1272–1283.

Muldoon, R. and M. Weisberg (2011). Robustness and idealization in models of cognitive labor. *Synthese 183*(2), 161–174.

NIH (2014). Success rates - NIH research portfolio online reporting tools (RePORT). `http://report.nih.gov/success_rates/`, Accessed 11 July 2014.

O'Malley, M. A., K. C. Elliott, C. Haufe, and R. M. Burian (2009). Philosophies of funding. *Cell 138*(4), 611–615.

Polanyi, M. (1962). The republic of science: Its political and economic theory. *Minerva 1*, 54–73.

Strevens, M. (2003). The role of the priority rule in science. *The journal of philosophy 100*(2), 55–79.

Weisberg, M. and R. Muldoon (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of science 76*(2), 225–252.

Wilkins, J. S. (2008). The adaptive landscape of science. *Biology and philosophy 23*(5), 659–671.

Zollman, K. J. (2010). The epistemic benefit of transient diversity. *Erkenntnis 72*(1), 17–35.

How To Make Selective Realism More Selective (and More Realist Too)

Massimiliano Badino
Massachusetts Institute of Technology — Universitat Autonoma de Barcelona

Abstract

Selective realism is the thesis that some wisely chosen theoretical posits are essential to science and can therefore be considered as true or approximately true. How to choose them wisely, however, is a matter of fierce contention. Generally speaking, we should favor posits that are effectively deployed in successful prediction. In this paper I propose a refinement of the notion of deployment and I argue that selective realism can be extended to include the analysis of how theoretical posits are actually deployed in symbolic practices.

1. Introduction

Among the several forms of realism, the so-called selective realism (SelRealism) is arguably the one that engages history of science more seriously. The driving idea of SelRealism is that, although theories as wholes are false and doomed to be abandoned, it is possible to select a certain number of theoretical posits (TPs) that are likely to be maintained in future theories and are therefore true or approximately true. How to determine these TPs is *partly* an empirical question—and this explains the historical character of the SelRealism program—but it cannot be *merely* an empirical question lest one end up in post-hoc rationalizations. A central issue of

SelRealism, hence, is how to specify criteria to properly conceptualize the TPs on which one should place one's realist commitment.

In this paper, I argue that contemporary approaches to SelRealism have neglected an important element related to the way in which theoretical claims are deployed in scientific theories (Section 2). In Section 3, I propose a refinement of SelRealism based on the distinction between deploying a TP fundamentally and deploying it in a non-accidental fashion. I use the concept of symbolic practices to articulate this distinction. Finally, in Section 4, I clarify my points by discussing the early development of perturbation theory.

2. Selective Realism: Theory and Practice

The upholders of SelRealism cherish two fundamental ambitions. First and foremost, they aim at making a good use of the so-called no-miracles argument (NMA) according to which one can justifiably infer the truth (or the approximate truth) of a successful theory, because, otherwise, the success would remained inexplicable. The NMA is considered to be the strongest support to realisms of any sort (Musgrave 1988; Psillos 1999, 68-94). A challenging objection to the NMA is the pessimistic meta-induction (PMI) originally formulated by Larry Laudan. According to this argument, the success of a theory is never a sufficient reason to infer even its approximate truth because history of science is replete with examples of very successful theories that wound up overthrown at some later stage. As it is likely the case that our most successful theories will suffer the same fate in the future, one has

to conclude that the realist commitment is not justified (Laudan 1981). Among the several responses to the PMI, one consists in noticing that the failures of past theories, in fact, did not depend on those TPs that lead them to success. In other words, granted Laudan's point that successful past theories are false as wholes, it can still be argued that the constituents of those theories that were responsible for their empirical success have been retained in our current science. Thus, the realist needs only to shift her commitment from theories as wholes to those enduring TPs that, being essential for success, can be justifiably believed to be true or approximately true.

The next question is, of course, how to determine those TPs. Thus, the second ambition of the upholders of SelRealism is to solve the problem of selectivity in some principled way and so beat the PMI. In one of the first instantiations of SelRealism, Philip Kitcher argued that one must "distinguish between those parts of theory that are genuinely used in the success and those that are idle wheels" (Kitcher 1993, 143). The point of this distinction is that credit for the success of a theory should be due only to those TPs that effectively contribute to it. Elaborating on Kitcher's intuition, one can argue that the program of SelRealism is based on two major conditions:

(S) Success condition: the selection of the important TPs must hinge on their relation with some significant success of the theory.

(D) Deployment condition: one must select those TPs that were effectively used in scoring that success.

Let me briefly comment on these two conditions. While (S) is now a realist trademark, the deployment condition (D) is what sets apart SelRealism from other forms of realism, such as structural realism, also engaged in picking out enduring elements of scientific theories (Worrall 1989; Chakravartty 2011). It is also important to notice that (S) and (D) are independent conditions. Firstly, (S) refers to a relation between the selected TP and empirical success, while (D) refers to a relation between the TP and the rest of the theory. Secondly, either condition can be satisfied separately. (D) has been added precisely to avoid those cases in which idle TPs are involved in empirical success and, obviously, there are scores of examples of TPs used by theories which however never led to any success. It follows that, while (S) is supposed to meet the first ambition of SelRealism, the second ambition, to block the PMI, is on (D).

So much for SelRealism in theory. Let us now examine how this program has been carried out in practice. One of the first philosophers to seriously elaborate on Kitcher's suggestion was Stathis Psillos. His criterion for selecting TPs works in the following way (Psillos 1999, 110). Let us assume that a certain successful prediction *P* can be obtained by combining the TPs *H*, *H*' and the auxiliaries *A*.[1] According to

[1] For virtually all writers, empirical success means "successful prediction". David Harker has leveled important criticisms against this tendency to interpret success in terms of individual predictions and has suggested that success should be understood as progress, i.e. in terms of the improvements a theory makes with respect to its predecessors (Harker 2008, 2013).

Psillos, the TP *H* is essential to success *P* and should be considered true or approximately true if and only if:

(1) *H*' and *A* alone do not lead to *P*.

(2) There is no alternative *H** to *H* such that:

(a) *H** is consistent with *H*' and *A*;

(b) *H**, *H*', and *A* lead to P;

(c) *H** is not *ad hoc* or otherwise purposefully concocted to lead to *P*.

This criterion is the bedrock of Psillos's *divide et impera* strategy. The driving intuition behind it is to capture the *indispensability* of *H*: we should place our realist commitment upon those TPs without which empirical success cannot be obtained. However, Tim Lyons has cogently argued that Psillos's criterion fails to characterize indispensability (Lyons 2006). The indispensability of *H* should be ensured by condition (2), which states, in brief, that *H* cannot be replaced by any other TP. But, Lyons notices, "there will always be other hypotheses, albeit some that we find very unappealing, from which any given prediction can be derived" (Lyons 2006, 540). More importantly, Lyons argues, Psillos's criterion is not even an effective means for credit attribution, because it does not tell us much about how *H* contributes to the empirical success *P*. In particular, condition (2) has no relevance whatsoever for *H*'s specific contribution, because it only concerns conceivable alternatives to *H*, alternatives that, if *H* is at hand, nobody would even bother to explore. Lyons

perceptively stresses that the problem with Psillos's criterion boils down to the fact that it obliterates condition (D): "by introducing his criterion, [Psillos] has discarded the central idea of deployment realism—introduced by Kitcher and seemingly advocated by Psillos himself" (Lyons 2006, 541). It is interesting to note that, by dropping condition (D), Psillos's position becomes vulnerable to another form of PMI. One could think of getting around of Lyons's first objection by arguing that, even though an alternative to *H* is always conceivable, *at the present state* of our knowledge it is not, therefore the objection is empty. In other words, one could inject the time factor in Psillos's criterion and make it a statement of our actual best knowledge. But then the PMI crops up again, because history shows that there is no guarantee that what is indispensable today will be so tomorrow. The whole point of the PMI is that there is nothing special in our knowledge as far as it is considered *present*, because there have been a lot of *present knowledges* that have been blissfully abandoned. This is why one needs condition (D): what makes our present knowledge so special is not its happening at a certain time, but its having gone through a certain *process*, i.e., a form of deployment. The fact that our present knowledge has been deployed at lengths and it is still with us constitutes a reason to believe that it is true or approximately true.

3. Deconstructing Deployment

Having grasped that the flaw in Psillos's criterion is the dropping of the deployment condition, Lyons suggests to run to the other end of the spectrum and to inflate

dramatically the notion of deployment. His "responsibility model" consists in discarding selectivity altogether and in considering responsible for the empirical success of a theory each and every element that was originally deployed: "credit will have to be attributed to all responsible constituents, including mere heuristics (such as mystical beliefs), weak analogies, mistaken calculations, logically invalid reasoning etc." (Lyons 2006, 543). Clearly, Lyons's proposal amounts to a crack-up of the entire SelRealism program. But, more importantly, I do not think that the responsibility model captures the correct significance of (D). As my previous considerations about the PMI show, the deployment condition is not merely supposed to tell us that a TP has been effectively used in obtaining empirical success (as opposed to be *dispensable*), but also that it has been robustly so (as opposed to be merely *accidental*). What makes it plausible that a TP will still play a role in future theories is the fact that its importance for empirical success has been tested by extensive and repeated deployment. It is therefore clear that there are two ideas nested in the deployment condition. One is the idea, captured by Psillos's criterion, that significant TPs must play a fundamental role in success in order to distinguish them from idle hypotheses; the other is the idea that the deployment of a TP must ensure that its success is not accidental. These are two distinct ideas. It might happen, for example, that a TP plays an essential role in deriving a prediction in virtue of fortuitous factors cancellation or other favorable circumstances. So, while an *intensive deployment* ensure the *fundamentality* of a TP, an *extensive deployment* founds its *robustness*. Both fundamentality and robustness are ways to articulate the complex relation between a

TP and the rest of the theory, or at least some parts of the theory (more on this in a bit). Further, while fundamentality is an atemporal articulation of this relation,[2] robustness concerns precisely the temporal dimension of the deployment condition that escaped Lyons's analysis: robustness, as we shall see below, is achieved over time.

In order to clarify the distinction between fundamentality and robustness, I introduce the notion of *symbolic practices*. By symbolic practices I mean all the methods customarily used in science to manipulate symbols.[3] These include, but are not limited to, mathematical methods, formal tools, approximations procedures, models, heuristics, solution tricks, and any sort of way by which one can transform a symbolic expression into another symbolic expression. Symbolic practices are the set of methods adopted by a theory to "put to work" a certain TP or, in other words, to deploy it in order to set problems and to interpret solutions. By using the concept of symbolic practices, one can reformulate the two ideas of the deployment condition in the following way:

---

[2] Of course the fundamentality of a TP can change over time because it can become more or less fundamentally used. However, the relation in itself does not concern this change.

[3] My discussion is especially tailored on the case of mathematical physics. I do not exclude, however, that it can be suitably extended to other branches of science by taking an appropriately enlarged notion of symbolic practices.

(F) Fundamentality: A TP must be *embedded* in a set of symbolic practices that lead to empirical success.

(R) Robustness: The symbolic practices adopted to deploy the TP must be *reliable*.

Let us begin with (F). This idea hinges on the "embeddedness" of a TP into a set of symbolic practices. An empirical success, a successful prediction or an explanation, is obtained by starting with one TP—or, better, its symbolic codification—and by deriving from it the phenomena to be treated by means of suitable manipulations. In their analysis of the path from TP to success, philosophers usually disregard the epistemic role played by symbolic manipulations of TPs. But if we neglect this important factor of the process of predicting/explaining, we are left with no other option than characterizing fundamentality as a relation between TPs, i.e., a 'Psillosian' criterion and then a 'Lyonsnesque' argument can easily prove that this falls short of providing a satisfactory notion of fundamentality. In my proposal, fundamentality is rather a relation between TP and the symbolic practices adopted to transform and manipulate it. Although intuitively clear enough, the concept of embededdness admittedly needs further philosophical analysis. In Section 4, I provide a historical example to clarify what it means for a TP to be embedded into a set of symbolic practices.

Before discussing the example, however, I need to analyze briefly the idea of robustness. Condition (R) states that reliability, and hence robustness, is a property of the symbolic practices themselves. In other words, and this is the central point, a TP

can be made more robust by means of *historically and rationally describable strategies* conceived to enhance the reliability of symbolic practices adopted to put it to work. One way to appreciate this point is to notice that the concept of reliability has three main components. First, there is an *empirical component*, that is its connection with success. It is expected that reliable symbolic practices have led and will lead to empirical success. This is unsurprising, because it is still part of the relation between (D) and the NMA. Second, there is a *conceptual component*: reliable symbolic practices allow us to distinguish between real facts of nature and artifacts. This is the component that accounts for the non-accidentality of success and it depends on the adoption of strategies to enhance reliability. Applying symbolic practices to multiple cases, relating them with other, better understood, sets of practices (e.g., by showing structure similarities), generalizing solution methods, simplifying computation procedures, introducing redundant check routines, improving the symbolic notation, multiplying proof procedures are just a few examples of strategies used to ensure that the result of symbolic manipulation is a real information and not an artifact generated by the practice itself.[4] Finally, there is a *historical component*. As I said above, deployment is a process extended over time. When are we justified to consider a result as reliable? This is an agent- and a context-dependent component of reliability.

[4] This component of the concept of reliability is closely connected with the usual notion of robustness (see, e.g., (Soler et al. 2012) for an overview). Indeed, robustness has to do with the multiplications of methods of check and control as a way to distinguish what is real and what is fabricated by practices.

I submit that this component can be clarified in terms of *control*. We develop theories because we need to manipulate symbols in order to make predictions and explanations. It is reasonable to state that an agent considers reliable a theory when she has control on it, when she knows how to do things, where the theory can be applied, to what extent, what kind of information she can obtain, what kind of epistemic risks are involved in it, how to improve progressively the performance and a lot of other things related to the general idea of knowing what is going on. Thus, reliability can change over time in virtue of new information and further inquiry. This component accounts for the fact that science is an ongoing human endeavor.

To sum up, I propose to extend SelRealism in the following way:

(SelRealism+) We are entitled to consider the TP *H* as true or approximately true at time *t* if and only if:

1. *H* is embedded into a set of symbolic practices *S*
2. *S* is reliable
3. *H* and *S* lead to significant success

This is a more selective version of SelRealism, because the philosophical and historiographical program stemming from it extends the inquiry to the strategies adopted to improve the reliability of symbolic practices and the contingent conditions for control. As stated in condition 3, the units of analysis of SelRealism+ are TPs-*cum*-

practices rather than TPs only. In the following section, I provide an example of what I mean by intensive and extensive deployment.

4. The Coming of Age of Perturbation Theory

The *Principia Mathematica* are a supreme example of how to embed a TP, in this case the gravitational law, into a set of symbolic practices.[5] However, Newton's mainly geometrical methods were fantastically complicated and notoriously difficult to master. A significant breakthrough in what came to be called celestial mechanics happened in the mid-1740s, when Leonhard Euler laid down the foundations of analytical perturbation theory. Euler made a number of decisive steps forward. First, he used the gravitational law to formulate general equations of motion for celestial problems. Second, he introduced the use of trigonometric series to construct approximate solutions. The use of these series also depended crucially on the gravitational law, because it satisfied the assumption that planetary orbits, even under perturbations, can be represented by a combination of periodic functions. Finally he introduced manipulation practices such as the method of the variation of

[5] In what follows, I consider perturbation theory as the set of practices conceived to put to work the gravitational law. It must be noted that other TPs were involved (e.g., Newton's laws of dynamics) and that the gravitational law can be decomposed in further assumptions such as the action-at-a-distance, the instantaneous propagation and so forth. These considerations affect the level of detail of my example, but not the structure of my argument.

constants and the method of successive approximations to solve the equations of motion. Perturbation theory is therefore a clear example of a set of symbolic practices conceived to cast a TP into a manipulable form and to applied it to specific problems.

For the purpose of this paper, I distinguish two phases in the early history of perturbation theory. The first phase goes roughly from the mid-1740s to the mid-1760s and it concerns the cause of numerous astronomical anomalies. Newton had left behind a few conundrums that even his genius was unable to unravel. The most conspicuous of these problems was the precession of the Lunar apogee. Newton's Lunar theory, elaborated in Book I and III of the *Principia* only managed to obtain half of the observed value. In the 1740s, there were two approaches to the issue of the Lunar apogee. The analytical approach adopted the gravitational law, or a slightly modified form of it, and tried to calculate the observed precession by analytical methods only. The physical approach supposed that the observed anomalies could be due to material causes such as a resisting medium or interplanetary vortices. It is important to realize that these approaches were compatible. Euler himself supported both the resisting medium hypothesis and the analytical approach and occasionally also proposed the use of vortices (letter to Clairaut, 30 September 1747). For several years, the best mathematicians of Europe struggled with the riddle of the Lunar apogee (Bodenmann 2010) until, on 21 January 1749, Alexis Clairaut showed that if one pushes the approximation to the second order of the perturbation, some terms that are negligible at the first order become sizable and generate the missing half of the precession (Clairaut 1752).

Clairaut's success was surely an impressive breakthrough, but what made it so impactful was not the brute fact that gravitational law had eventually led to a successful explanation. Physical hypotheses such as vortices and resisting medium also provided an explanation of the observed precession. The crucial difference lies in the fact that the gravitational law could be fully integrated with the analytical practices and then manipulated to provide suitable symbolic expressions of the precession of the apogee. That did not happen with the physical hypotheses, although not for lack of trying. Euler, for instance, tried hard to integrate the hypothesis of the resisting medium in perturbation theory, but the ensuing equations of motion were simply unmanageable (Euler 1747). Clairaut's success is eminently a story of intensive use of the gravitational law: he managed to integrate it with a set of symbolic practices and to accommodate effectively the observations.

Clairaut's feat did not close the debate on the gravitational law, tough. His calculations used many case-based assumptions, simplifications, and shortcuts and its straightforward extension to more complex cases, such as the behavior of Jupiter and Saturn, was doubtful to say the least. But there was also a deeper problem. At some point in his analysis, Clairaut obtained an "arc of circle", i.e., a trigonometric function multiplied by time. Such terms are obviously unbounded and hence make the whole trigonometric series diverge. Clairaut got rid of it by ad-hoc assumptions, but the status of these unbounded terms remained unclear: they could represent an artifact of the theory, a limitation of its predictive power or even a dynamical instability of the system.

Soon, the problem of the arcs of circle become more troublesome. Euler found the same terms in his analysis of the motion of Jupiter and Saturn and in 1766 Lagrange proved that they are actually a necessary consequence of the method of successive approximations applied to astronomical problems (Lagrange 1766). Thus, in the mid-1760s, perturbation theory appeared to be a fragile set of practices which had scored some important success, but was still marred with problems of unreliability under certain conditions. From the late 1760s onwards, the issue of improving the robustness of perturbation theory became a central preoccupation of the leading mathematicians interested in physical astronomy.

There were two programs inspired by this issue. On the one hand, Lagrange tried to improve the reliability of perturbation methods *as a mathematical theory*. He carried out this project by means of multiple strategies: (1) enhancing the relation between perturbation theory and other branches of mathematics (e.g., potential theory); (2) elaborating arguments to extract information from the equations of motion without solving them (e.g., by using integrals of motion); (3) improving methods to simplify the solution procedure (e.g., Lagrange's coordinates); (4) introducing new symbolic codifications to manipulate the equations of motion (e.g., the perturbing function); (5) making the notation less cumbersome (Lagrange's coefficients). Around the same years, Laplace was also working to improve the reliability of perturbation theory, but his program adopted a different approach. He concentrated on methods to make perturbation theory a more reliable *problem-solving tool*. He developed his own method to eliminate the arcs of circle—which was based on the recalculation of the

integration constants—he imported probability theory and the equations of condition to deal with astronomical observations and devised several strategies to identify in concrete cases those elements of the equations of motion that were likely to produce sizable perturbation terms at higher order. Both Lagrange's and Laplace's programs scored their own successes. In the early 1780s, Lagrange proved a very general result of stability according to which the three more important orbital elements (mean motion, eccentricity, and inclination) are invariable or bounded (Lagrange 1781). Laplace, on his part, explained the decades-long problems of the anomaly in the motion of Jupiter and Saturn as well as the secular acceleration of the Moon (Laplace 1785, 1787; Wilson 1985).

5. Conclusions

In several places, Kyle Stanford has argued that any selection of enduring TPs is ultimately ungrounded and, consequently, the entire SelRealism program is unviable (Stanford 2003, 2006). In his view, there are two possible ways to select essential TPs. The first way is to trust scientists when they say that a certain posit is fundamental. However, neither commonsense, nor, more importantly, historical records support the hypothesis that scientists' take on this matter is or should be particularly reliable. The other option is to wait and see: when a theory is superseded, one can check which TPs have survived. The reason why a selective realist cannot go with this option, however, has been summarized effectively by Peter Vickers:

> If we cannot identify the working posits of a theory until it has been superseded by some other theory, then realism is no longer about identifying what we ought to believe to be true: one is always waiting for the next theory to come along to tell us which parts of our current theory are working posits. (Vickers 2013, 207)

From this, Stanford concludes that SelRealism without prospectively applicable selectivity criteria is empty and should be replaced by a more modest form of realism. But Stanford's wait-and-see stance is neither necessary nor sufficient to do the job it is supposed to do, i.e., to pick out essential TPs. It is not sufficient because there is no guarantee that the TPs survived one theory change will survive the next ones. It is not necessary because we do not need the next theory to form reasonable judgements about essential TPs. As I have shown above, science provides a variety of strategies to improve the reliability of the TP-*cum*-practices and hence good reasons to believe, *within the actual theory*, that a certain TP intensively and extensively deployed is in fact essential.

From this perspective, Stanford's argument simply sets the epistemic bar too high. By stating that the essentiality of a TP can be adjudicated only from the vantage point of the superseding theory, he implicitly challenges the realist to provide a "superselection rule" able to capture the whole history of science, a task that the realist is neither willing, nor actually requested to accomplish. By contrast, the historical and philosophical program of SelRealism+ moves from the conviction that TPs and symbolic practices follow a dynamics able to filter out inessential

components. Consequently, SelRealism+ is committed to historically identify and philosophically analyze this dynamics and to trace the genealogy of our theories in terms of the processes of codification, manipulation, and stabilization of TPs. Ultimately, this program aims at producing new and interesting historical narratives of theory change. It remains true that the strategies making up the theoretical dynamics only provide good reasons to allocate the realist commitment. It might happen that the judgement on the reliability of the TPs-*cum*-practices change over time in virtue of further inquiry or new information. This fact, as stated above, follows from the fallibility of science as a human endeavor and, as such, should not trouble the realist.

Acknowledgements

The research for this paper has been supported by the Marie Sklodowska-Curie Actions, grant no. PIOF-GA-2013-623436.

References

Bodenmann, Siegfried. 2010. "The 18th Century Battle over Lunar Motion." *Physics Today* no. 63:27-32.

Chakravartty, Anja. *Scientific Realism* 2011 [cited 4 February 2015. Available from http://plato.stanford.edu/entries/scientific-realism/.

Clairaut, Alexis. 1752. "De l'orbite de la lune, en ne negligeant pas les quarrés des quantités de meme ordre que les forces perturbatrices." *Memoire de L'Academie Royale des Sciences*:421-440.

Euler, Leonhard. 1747. "Recherches sur le mouvement des corps cèlestes en général." In *Opera Omnia*, 1-44. Leipzig: Teubner.

Harker, David. 2008. "On the Predilections for Predictions." *British Journal for the Philosophy of Science* no. 59:429-453.

———. 2013. "How To Split a Theory: Defending Selective Realism and Convergence without Proximity." *British Journal for the Philosophy of Science* no. 64:79-106.

Kitcher, Philip. 1993. *The Advancement of Science*. Oxford: Oxford University Press.

Lagrange, Joseph Louis. 1766. "Solution de différents problèmes de calcul intégral." In *Œuvres de Lagrange*, edited by Jean A. Serret, 609-668. Paris: Gauthier-Villars.

———. 1781. "Théorie des variations périodiques (Premiére partie contentant les formules générales de ces variations." In *Œuvres de Lagrange*, edited by Jean A. Serret, 347-377. Paris: Gauthier-Villars.

Laplace, Pierre S. 1785. "Théorie de Jupiter et de Saturne." In *Œuvres de Laplace*, 95-239. Paris: Gauthier-Villars.

———. 1787. "Memoire sur les Variations seculaires des Orbites des Planetes." In *Œuvres de Laplace*, 295-306. Paris: Gauthier-Villars.

Laudan, Larry. 1981. "A Confutation of Convergent Realism." *Philosophy of Science* no. 48:19-49.

Lyons, Timothy D. 2006. "Scientific Realism and the Stratagema de Divide et Impera." *British Journal for the Philosophy of Science* no. 57:537-560.

Musgrave, Alan. 1988. "The Ultimate Argument for Scientific Realism." In *Relativism and Realism in Science*, edited by Robert Nola, 229-252. Dordrecht: Kluwer.

Psillos, Stathis. 1999. Scientific Realism: How Science Tracks Truth. London: Routledge.

Soler, Lena, Emiliano Trizio, Thomas Nickles, and William C. Wimsatt. 2012. *Characterizing the Robustness of Science*, *Boston Studies in the Philosophy of Science*. Dordrecht: Springer.

Stanford, P. Kyle. 2003. "No Refuge for Realism: Selective Confirmation and the History of Science." *Philosophy of Science* no. 70 (913-925).

———. 2006. Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives. Oxford: Oxford University Press.

Vickers, Peter. 2013. "A Confrontation of Convergent Realism." *Philosophy of Science* no. 80:189-211.

Wilson, Curtis A. 1985. "The Great Inequality of Jupiter and Saturn: from Kepler to Laplace." *Archive for History of Exact Sciences* no. 33:15-290.

Worrall, John. 1989. "Structural Realism: The Best of Both Worlds?" In *Philosophy of Science*, edited by David Papineau, 139-165. Oxford: Oxford University Press.

Duhemian good sense and agent reliabilism

Famously, according to Duhem a hypothesis can never be experimentally tested in isolation, but only along with the entire theoretical scaffolding it comes with. So in the face of disagreement between theory and experiment, it is impossible to point out which hypotheses in the theory are flawed. A big question for Duhem was, how does the physicist act in such a situation of underdetermination? Which hypotheses does s/he discard, and which one(s) does s/he retain? Duhem's response was that the physicist possesses an intuitive "good sense" that directs this choice. Although good sense does not provide a rigorous, rule-based template for theory choice[1], it allows scientists to weigh evidence and be "fair and impartial" (Duhem, 218) in theory choice.

Recently, there has been much interest in drawing parallels between Duhem's good sense and ideas in virtue epistemology (VE). VE emerged in the 1980s as an approach to epistemology based on virtue ethics. In the words of Greco (2004): "Just as virtue theories in ethics try to understand the normative properties of actions in terms of the normative properties of moral agents, virtue epistemology tries to understand the normative properties of beliefs in terms of the normative properties of cognitive agents." A virtue epistemological reading of good sense as first advanced by David Stump (2007) is based on the idea that Duhem too emphasized the normative properties of the scientist qua cognitive agent and took them as a basis for legitimate scientific

[1] While "theory choice" today is generally understood in the context of contrastive underdetermination, Duhem was primarily concerned with the holist variety of underdetermination and advanced good sense in the context of the latter. But for the purposes of this paper the distinction will not matter, and I shall use "theory choice" to refer to underdetermination in general, as do all the authors I reference.

knowledge in the face of underdetermination of theory by evidence. Stump finds striking similarities particularly between Duhemian good sense and Linda Zagzebski's (1996) views of VE. Here, I discuss the views of Stump, Milena Ivanova (2010), and Abrol Fairweather (2012) in this regard and ultimately propose my own view in response which is an agent-reliabilist reading of Duhem's good sense.

Stump argues that Duhem conceived of good sense in a way that can today be understood as virtue theoretic. In particular, Stump finds similarities between good sense and ideas of VE put forward by Zagzebski (1996). As Stump tells us, Zagzebski argued that justified belief comes from a "cluster of intellectual virtues in the same way that the rightness of an act can be defined in terms of moral virtue in ethical theory"(Stump, 151). Stump argues that Duhem's good sense nicely fits in with these ideas. Good sense depends on the scientist, the cognitive agent, being "virtuous": s/he has to be, in the words of Duhem quoting Claude Bernard, a "faithful and impartial judge". Stump further provides another illuminating quote from Duhem from his lectures on German science:

> "In the realm of every science, but more particularly in the realm of history, the pursuit of the truth not only requires intellectual abilities, but also calls for moral qualities: rectitude, probity, detachment from all interest and all passions. (Duhem, 1991b, p. 43)" (Stump, p. 152).

Stump notes that some of the epistemic virtues put forward by Zagzebski include intellectual sobriety, impartiality and intellectual courage and the list fits very well with Duhem's. Yet another striking similarity between Zagzebski and Duhem according to Stump is that they both appeal to non-rule-governed epistemology. Zagzebski, in making a case for an

epistemology based on ethics, says, "The idea is that there can be no complete set of rules sufficient for giving a determinate answer to the question of what an agent should do in every situation of moral choice." (Stump, 152) Similarly, Duhem arrives at the idea of good sense when the rule-based epistemology of the physical method (i.e. strict agreement between theory and experiment) fails. As Stump says,

> "Holism threatens to make testing impossible, yet Duhem believes that scientific consensus will emerge. While the pure logic of the testing situation leaves theory choice open, good sense does not. Duhem claims that the history of science shows that while there is controversy in science, there is also closure of scientific debates." (Stump, 155)

Milena Ivanova (2010) has argued in response to Stump, that the latter is mistaken in drawing such close parallels between VE and Duhem's good sense. She raises two main objections: first, while VE is concerned with getting to the *truth* via epistemic virtues, for Duhem, physical theory only asymptotically approaches truth – truth here being the truth of a natural order, of the "real affinities" among things. Ivanova makes this point keeping in mind Duhem's view of a 'perfect theory' and the convergent nature of his realism: for Duhem, the aim of physical theory was to classify experimental laws, and a physical theory – one picked out by good sense in the face of underdetermination – constantly approached but never reached, a perfect theory which classified laws and their phenomena in exactly the way underlying metaphysical realities are really classified in nature. So her point is that while VE is concerned with getting to the truth, good sense doesn't help us with that. But as Ivanova herself points out,

> "Still, in response to this objection one can adopt the weaker thesis that even though natural

> classification may not reveal the truth about the unobservable, it will be true for the observable phenomena. Also, one may argue that it is legitimate to aim at a particular epistemic goal independently of whether this goal is achievable or not." (62)

I take her point here to be that both VE and good sense are after all in the business of truth-seeking even though attaining the truth may be impossible for with the latter.

Ivanova's more forceful objection has to do with epistemic justification. According to her whereas VE takes epistemic virtues to be *justifications* for beliefs, Duhem did not invoke the concept of good sense to *justify* belief in one theory over another. (To reiterate, Duhem did not have a full-blown metaphysical notion of truth of a theory – but worked with the surrogate idea of truth, that a right theory approaches a transcendental, natural classification.) Rather, she argues, good sense for Duhem was more a post hoc *explanation* of the physicist's choice: it explains the repeated success of theories at making novel predictions. According to Ivanova, what really justified belief in a theory for Duhem – i.e. the belief that it was approaching a natural classification – was the success of the theory in making correct novel predictions: She says that for Duhem, "[a scientist] is justified in believing that a theory is a natural classification only when some empirical evidence supports it or when the theory has become a 'prophet for us' (Duhem, 27), that is, when it has managed to make novel predictions." (Ivanova, 62). Here's Ivanova's argument broken down:

- Physical theory is a classification of laws.
- In a situation where we have a theory that contradicts experimental data and are left without any means within physics to decide what to do - whether to tweak parts of the theory to accommodate the available experimental data – and if so, which parts to tweak

– or to abandon it for another theory. Somehow in the end, the scientist decides which way to go.

- The "highest test" for physical theory is to ask it to make new and novel experimental predictions.
- When the theory succeeds it is justified – in that it is taken to approach a natural classification.
- Repeatedly, the scientist sees her/his choices made in the difficult situation of underdetermination emerging successful in such predictions.
- How does this happen? There must be some innate ability or virtue in the scientist that enables him to do this: good sense.

Thus according to Ivanova, good sense is an explanation of theory choice rather than a justification for it. Moreover, according to her, Duhem doesn't say anything about good sense as a method of science: he doesn't tell us *how* exactly it directs our choice. His account of how good sense comes about and works to direct theory choice is quite thin. For Ivanova, this further shows that Duhem did not introduce it as a justification but only as a post hoc explanation.

Abrol Fairweather (2012) has argued against Ivanova's above objection and has attempted a position on Duhemian good sense that is a hybrid of Ivanova's and Stump's views. Fairweather claims to draw upon an agent reliabilist VE to do this. Reliabilism in Alvin Goldman's words, "… as a distinctive approach to knowledge is restricted to theories that involve truth-promoting factors above and beyond the truth of the target proposition." (Goldman, 2011) Fairweather's argument is that good sense results in a *reliable* process. Since Duhem's

claim is that good sense has a great "track record" and always picks out a successful theory – i.e. a theory which inevitably *correctly* makes a novel prediction – good sense produces knowledge (which here in the Duhemian context, consists in taking a predictively successful theory to be approaching a natural classification) by a *reliable* process. Good sense is a 'truth-promoting factor' regardless of whether the theory it picks out ultimately succeeds in novel prediction or not. It is "tracking evidentially important features of theories" (Fairweather, 10) Fairweather claims that "If a belief P is the product of a reliable capacity or process this fact constitutes evidence in favor of P." This implies, "If the products of good sense reliably turn out to be supported by compelling new evidence, then being the product of good sense will be evidence for any theory with such a distinguished etiology." (Fairweather, 10) So, Fairweather says, it seems that "future evidence is not required to evidentially distinguish the theory chosen by good sense, because the reliability of good sense is itself evidence supporting that theory." (Fairweather, 10) While I agree that agent reliabilism is the best way to understand good sense, Fairweather does not seem to give an accurate interpretation of this reading. Although he claims to provide an agent reliabilist reading of good sense, he grounds the reliability of good sense in its track record and not in its own nature or the mind where it is born. This is antithetical to agent reliabilist VE which situates reliability in the cognitive character of the agent. So it seems that Fairweather's characterization is more along the lines of process reliabilism or simple reliabilism – according to which a belief is justified just in case it is formed via reliable processes – rather than agent reliabilism, and hence contrary to what he set out to do. His argument does not help situate good sense back into VE. Let us now turn to agent reliabilism in detail.

Greco and Agent Reliabilism: A Short Detour

As above, simple reliabilism is the view that a belief is justified just in case it is formed via reliable processes. Here the proportion of true beliefs the process results in, over time, measures reliability. Greco (1999) argues that simple reliabilism is insufficient for two reasons:

1. An agent might form a belief via fleeting or strange processes: Greco starts by noting that "Reliabilism must somehow restrict the kind of reliable process that is able to ground knowledge, so as to rule out processes that are strange or fleeting." (Greco, 286) As an example of such processes, Greco discusses Platinga's "The case of the epistemically serendipitous lesion" where an agent has a rare kind of a brain lesion, one that makes her believe that she has a brain lesion. There is no evidence for the lesion: there no symptoms, no testimony etc.; in fact there might even be a lot of evidence *against* it. But the agent is unable to take account of this (lack of) evidence due to the lesion. The relevant cognitive process here must no doubt be deemed very reliable, but we would not want to take the resulting belief as justified.
2. Process reliabilism doesn't guarantee that the agent has a subjective justification of her belief. Greco says,

> "[there] is a powerful intuition that knowledge does require that the knower have some kind of sensitivity to the reliability of her evidence. Sometimes this intuition is expressed by insisting that knowledge requires subjective justification. It is not enough that one's belief is formed in a way that is objectively reliable; one's belief must be formed in a way that is subjectively appropriate as well." (285)

Greco's solution to the above problems is agent reliabilism. According to agent reliabilism, reliability is shifted from the belief-forming process to the qualities of the agent's

mind:

> "Relevant to present purposes is Sosa's suggestion for a restriction on reliable cognitive processes; it is those processes that have their bases in the stable and successful dispositions of the believer that are relevant for knowledge and justification. Just as the moral rightness of an action can be understood in terms of the stable dispositions or character of the moral agent, the epistemic rightness of a belief can be understood in terms of the intellectual character of the cognizer." (Greco, 287)

Following Sosa's views, Greco proposes that "knowledge and justified belief are grounded in stable and reliable cognitive character."(Greco, 287) Accordingly, "We may now explicitly revise simple reliabilism as follows: A belief p has positive epistemic status for a person S just in case S's believing p results from stable and reliable dispositions that make up S's cognitive character." (Greco, 287) Hence we see that reliability now has little to do with the truth of the resultant belief(s) but rather with the cognitive character of the agent.

Greco proceeds to show how agent reliabilism also solves the problem of subjective justification:

> VJ: "A belief p is subjectively justified for a person S (in the sense relevant for having knowledge) if and only if S's believing p is grounded in the cognitive dispositions that S manifests when S is thinking conscientiously." (289)

By "thinking conscientiously", Greco clarifies that he does not mean thinking with the purpose of finding truth, but rather the "usual state that people are in as a kind of a default mode – the state of trying to form beliefs accurately." Greco contrasts this with epistemic "vices" such as trying to comfort oneself or trying to seek attention. Lastly, Greco points out that agent reliabilism reverses the "usual direction of analysis between virtuous character and justified

belief". While non virtue theoretic epistemologies understand virtues in terms of justified belief, here justified belief is being cached out in terms of virtues of the cognizer. "Virtuous belief is associated with the dispositions a person manifests when she is sincerely trying to believe what is true", and "The dispositions that a person manifests when she is thinking conscientiously are stable properties of her character, and are therefore in an important sense hers." (Greco, 290) Therefore, a belief formed this way will be subjectively appropriate.

Back to Duhem

Duhem's views seem to exhibit all the features of agent reliabilism discussed above. In addition to the features of good sense and the physicist qua cognitive agent discussed so far I want to draw the reader's attention to Duhem's characterization of the different kinds of minds. For Duhem, the "strong and the narrow" mind is one capable of ordering and organizing laws and hypotheses into theories, and the "supple" mind or the "mind with finesse" – one capable of grasping a wide range of objects and at the same time able to group them logically – is the mind that produces good sense. This certainly seems to talk of "stable dispositions" in Greco's sense of the term, that reflect the "cognitive character" of the scientist. Duhem takes pains to carefully describe the mind of the physicist and discuss beliefs and attitudes *in terms of* cognitive character traits and not the other way round. i.e. Duhem talks of legitimacy of beliefs in terms of cognitive character traits; he does not talk of the traits or "epistemic virtues" so to speak, in terms of the validity of beliefs. For instance, he says about those not interested in seeing a unified system of classification erected, "Only those who affect a hatred of intellectual strength were mistaken to the extent of taking the scaffolding for a completed building." (Duhem, 103) There are several such instances where Duhem turns traditional non virtue-theoretic epistemology on its head and makes cognitive character traits basic. Now it remains to be seen if we can defend a view of

justification from good sense that goes with Greco's account. If we are successful in this, Ivanova's position will be untenable. Before going there though, let us return to Fairweather for a moment.

In addition to the argument from reliabilism, Fairweather advances another argument against Ivanova's "deflation of good sense": the position that good sense does not lend any epistemic strength or any justification to the chosen theory. The argument is that if good sense were indeed merely explanatory and post hoc as Ivanova claims, and not justificatory, then we are free to imagine a case where good sense doesn't intervene at all. After all, if good sense explains theory choice and there is no choice being made – i.e. no explanandum - we don't need an explanation. So let us suppose that we don't make any choice and just wait for a future novel prediction to make a choice and justify it. This might not be the most efficient way to choose a theory, but let us assume we do this nevertheless – for according to Fairweather, Ivanova's objection should imply the possibility of this solution. Fairweather rightly points out that in this situation we *might* again end up with an underdetermination: what if all competing theories pass the novel prediction test? Therefore, Fairweather argues, good sense must play an important epistemic role above mere explanation, in the face of such a "second level" underdetermiantion. But he goes further than that and says that without it, we *would* never end up with a determinate choice, even with new confirming evidence. What Fairweather is ignoring here is that future evidence *could* pick out a theory, however small the probability. It is possible that when all the options resulting from underdetermination are asked to make a novel prediction, only one succeeds, hence obviating the need for any further theory revision. But the important point is that good sense enters the scene even before such an attempt to single out a theory based on novel prediction. So the merit of good sense in my view does not lie in the inability of novel

predictions to single out a theory. It is more fundamental than that. But reasons for meriting good sense apart, let us again look at Fairweather's take on *what* the merit of good sense is.

According to Fairweather, good sense confers *uniqueness* to a theory (which, according to him, no future evidence can confer). But after good sense has uniquely picked out a theory, it is a successful novel prediction that counts as evidence in favor of the chosen theory. Fairweather makes the following interesting observation that follows from such a reading of good sense:

> "This shows an interesting fact that new evidence in favor of a theory gives it a different epistemic standing depending on whether we are considering it alongside or independent of meaningful rivals. In the former case, new confirming evidence does not make a theory the determinate choice with fundamental epistemic standing. In the latter case, that same evidence determines theory choice and confers fundamental epistemic standing." (Fairweather, 13)

So there are two "epistemic values and epistemic standings": uniqueness, which comes from good sense, and clinching evidential support from a successful novel prediction. This way, good sense alone does not confer "fundamental epistemic standing", and evidence alone cannot confer uniqueness. This account which recognizes an important epistemic role for both good sense and new evidence, Fairweather calls the "hybrid reading".

My own view is that while Fairweather is right in that good sense plays a key epistemic role unlike what Ivanova says, we can go back full circle to Stump and have a proper virtue epistemological – specifically agent reliabilist – reading of good sense. I contend that good sense confers not just uniqueness, but actually does determine theory choice, also providing (an agent-

reliabilist) justification. Good sense doesn't simply pick one and put the rest "out of the running". It is not just something that prevents the proliferation of acceptable theories obtained by tweaking different parts of theories that don't agree with future experiment. Good sense provides a *basis* for the uniqueness. Just as with the problem of coming up with a realist interpretation of Duhem, this problem of the epistemic role of good sense is not easy either given the sometimes confusing nature of Duhem's claims. Nonetheless, I still think an agent-reliabilist VE reading of Duhem is possible and that Ivanova and Fairweather are mistaken.

Ivanova claims that good sense is only offered as a post hoc explanation of theory choice during underdetermination and not as a justification. I argue to the contrary. Ivanova's claim seems to be based on a purely externalist notion of justification. It seems to assume that there is one single concept of justification – specifically, externalist, evidential – and that good sense doesn't fit with it. But justification can be of many kinds. Duhem says we can "very properly decide" (Duhem, 217) between multiple theory choices using good sense. Further, he says good sense strongly "comes out in favor of" one of the choices – again implying that we are compelled to accept its judgment *even before* future experiment can ratify the choice. He goes on to say, "Pure logic is not the only rule for our judgments; certain opinions which do not fall under the hammer of contradiction are in any case perfectly unreasonable." (Duhem, 217) How do we understand such language? If an epistemic choice is proper, forceful, and reasonable, I don't see any reason we cannot properly construe it as being justified, in an *internalist* sense.

Further, Duhem does *not* introduce good sense as a merely post hoc explanation. He says, we can "properly decide" between the various options of theories using good sense. "Properly

decide" very much implies an active role for good sense *during* underdetermination. Duhem presents elaborate and careful characterizations of different kinds of minds and puts forward quite clearly, *normative* merits of cultivating/ possessing one kind of mind over the other as far as physics goes (the supple or the strong and narrow over the ample, broad and weak). Good sense is but a feature of the supple mind. It is not introduced all of a sudden as a new idea to just "save the (meta)phenomenon" of theory choice during underdetermination. It is a smooth and natural continuation of Duhem's views on the mind of the theorist, which he articulates way before he comes to this problem of underdetermination, in one of the early chapters in *Aim and Structure*. In fact, Duhem's view that physicists don't actually actively choose hypotheses at all, and that they "come to his mind" when his mind is ready to receive them, clearly reveals the agent reliabilist in Duhem.

Finally, Greco's account of agent reliabilist justification seems to lend itself to Duhem very well. Reliable cognitive character *justifies* beliefs it produces and further, it is subjectively justified: Duhem's virtuous scientist certainly "thinks conscientiously", following Duhem's instructions of shunning passions and interests, and so a belief, here the belief in the theory chosen, grounded in the cognitive dispositions, here good sense, he manifests when thinking like this – is subjectively justified. So we seem to have comfortably accommodated Duhem in a full-blown agent reliabilist reading.

But what about the textual evidence cited by Ivanova, which seems to say Duhem did not think good sense justified theory choice? Why does Duhem insist that despite good sense, it is a successful novel prediction that has the final word? Why does he, in the context of resolving

underdetermination say in as many words that the method of the physicist "is justified only by experiment"? I contend that throughout *Aim and Structure*, Duhem seems to have two distinct, non-intersecting epistemologies: one of physics, and one outside of physics – which we may call philosophy. Duhem was a physicist-philosopher. He frequently claims that although there are absolutely no epistemic resources *within* physics for us to believe that physical theory latches on to a natural underlying order, we are forced to believe so by various factors outside of physics, logic and reason. It is worth noting that Duhem cites Pascal as saying that we sometimes believe for 'reasons that reason does not know', both in the context of theories converging on to a natural classification as well as in that of good sense during underdetermination. About the former, he says: "The opinion is a legitimate one because it results from an innate feeling of ours which we cannot justify by purely logical considerations, but which we cannot stifle completely either." (Duhem, 102) Further:

> "No language is precise enough and flexible enough to define and formulate them; and yet, the truths which this common sense reveals are so clear and so certain that we cannot either mistake them or cast doubt on them; furthermore, all scientific clarity and certainty are a reflection of the clarity and an extension of the certainty of these common-sense truths." (Duhem, 104)

Since Duhem attributes good sense to similar patterns of thinking, we can associate his above assertions about the legitimacy of beliefs not borne out of logic, with good sense as well. Given Duhem's commitment to the moral goodness and the intellectual acuity of the supple, strong and narrow minds, it is very unlikely that he would think that epistemic ends justify the means (here, successful novel prediction justifying that which chose the theory, i.e. good sense). Reliabilism in fact expressly turns this around and say it is the means (by virtue of their

reliability) that justify the ends. So beliefs that arise from good sense are *justified* from an (internalist, deontological) agent reliabilist perspective. The justification Duhem talks about when he says that the methods of the physicist are justified by experiment should be when we are strictly within the context of physics: there it is Duhem qua physicist speaking. But from a broader, philosophical perspective, Duhem rather means, I think, that experiment *validates* the choice and confers *certainty* on it. But we can have justification without certainty, like in agent reliabilism. In simpler terms, the *reasons* for which the physicist chooses a theory are grounded in her good sense. However, the successful novel prediction will no doubt make the choice certain.

Thus, Ivanova is mistaken in arguing that good sense does not provide justification. Fairweather's hybrid reading is inadequate as well for it ignores the justification offered by a proper agent reliabilist reading of good sense. I argue that a proper agent reliabilism accommodates Duhem as a virtue epistemologist very well and shows us that good sense does offer justification for theory choice. Importantly, I have shown that it is certainly not a post hoc explanation but a part and parcel of Duhem's overall views on the mind of the physicist.

References

- Duhem, Pierre. (1954). *The aim and structure of physical theory*. Princeton: Princeton University Press.
- Fairweather, A. (2012) 'The Epistemic Value of Good Sense' *Studies in the History and Philosophy of Science* http://philpapers.org/archive/FAITEV.pdf

- Goldman, Alvin. 'Reliabilism', *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2011/entries/reliabilism/>.

- Greco, J. 1999. 'Agent reliabilism' in *Philosophical Perspectives* 13: 273-296.

- Ivanova, M. (2010). 'Pierre Duhem's good sense as a guide to theory choice'. *Studies in History and Philosophy of Science*, 41, 58–64.

- Stump, David. (2007). Pierre Duhem's virtue epistemology. *Studies in History and Philosophy of Science, 38, 149–159*.

There *Is* a Special Problem of Scientific Representation

(Word count: 4998)

Abstract: Callender and Cohen (2006) argue that there is no need for a special account of the constitution of scientific representation. I argue that scientific representation is communal and therefore deeply tied to the practice in which it is embedded. The communal nature is accounted for by *licensing*, the activities of scientific practice by which scientists establish a representation. A case study of the Lotka-Volterra model reveals how the licensure is a constitutive element of the representational relationship. Thus, any account of the constitution of scientific representation must account for licensing, meaning that there *is* a special problem of scientific representation.

1. Introduction

According to many philosophers of science, representation in scientific practice is different from representation in other disciplines, like art and language. This claim is denied by Craig Callender and Jonathan Cohen (2006), who argue that representation is the same across disciplines. In this paper, I will argue that their view leaves the communal nature of scientific representation unexplained. To explain why scientific representation is dependent upon practice, I will introduce the concept of licensing, in which the targets of representational vehicles are determined through various activities performed by scientists in accord with broader scientific practice. I will argue that licensure is a constitutive feature of representation in science, indicating that there *is* a special problem of scientific representation.

2. Callender and Cohen's View

On Callender and Cohen's evaluation, much of the literature on scientific representation has been "concerned with non-issues" (2006, 67). Specifically, they think there is no reason for philosophers of science to give a special account of the "constitution question:" "What constitutes the representational relation between a model and the world?" (2006, 68). In response to this question, they make a few observations. One is that it is "economical and natural to explain some types of representation in terms of other, more basic types of representation" (2006, 70). They also identify a general desire to have a consistent account of how "entities other than models—language, pictures, mental states, and so on—…represent the very same targets that models represent" (2006, 71). For these reasons, they suggest that

"scientific representation is just one more special case of derivative representation" (2006, 75). That is to say that the representational nature of scientific vehicles is explained in the same way that the representational nature of linguistic entities, artwork, etc. is explained. In each case, and in every practice, the representational nature in question will be reduced to a more fundamental representational entity. So, e.g., the representational nature of a word, a painting, and a scientific model will each be explained in terms of the representational nature of mental states.

On Callender and Cohen's view, representation is purely stipulative: "virtually anything can be stipulated to be a representational vehicle for the representation of virtually anything…" (2006, 74). Of course, it is not the case that *any* stipulated representation will actually be useful for scientific aims. Thus, they identify pragmatic constraints which delimit scientific representation. However, they make it quite clear that these constraints are delimiting *already-existing* representations. As such, the pragmatic constraints are not a part of an account of the constitution of representation itself: "the questions about the utility of these representational vehicles are questions about the pragmatics of things that are representational vehicles, not questions about their representational status per se" (2006, 75).

If Callender and Cohen are correct, then we are left rethinking a rather extensive literature on scientific representation which typically begins with the assumption that there *is*

something special about representation in science.[1] As one example among many, Mauricio Suárez (2004) defends an inferential conception of scientific representation. His account takes careful notice of the aims of scientific practice, noting that mere stipulation (what he calls "representational force") is insufficient for representation in science. To be a *scientific* representation, a vehicle must also permit surrogate reasoning which "allows competent and informed agents to draw specific inferences regarding [a target]" (2004, 773). If we accept Callender and Cohen's view, then Suárez's account and the many others like it do nothing more than identify some of the typical pragmatic strategies employed in delimiting representations for scientific uses (Callender and Cohen 2006, 78).

3. Private Reminiscence and Communal Representation

In order to show that the extensive literature on scientific representation has not been addressing a non-issue, I will need to show that there is a special problem of scientific representation, a feature unexplained by Callender and Cohen's account. I submit that the relevant feature in need of special explanation is the communal nature of scientific representation, that it inherently involves reference to the practice. To see why Callender and

[1] For more accounts which answer the constitution question in a distinct way, see the work of Ronald Giere (1988, 2004), Bas van Fraassen (1980, 2008), RIG Hughes (1997), Steven French, James Ladyman, and Otávio Bueno (French and Ladyman 1999; Bueno and French 2011), and Gabriele Contessa (2007). For an overview of these accounts of scientific representation among others, see Brandon Boesch (2015) and Mauricio Suárez (2015).

Cohen's view is unable to account for the communal nature of scientific representation, consider what I call 'reminiscence', a representational relationship which lacks the same communal feature. It is defined schematically as the following:[2]

> Some X is reminiscent of some Y for some agent A provided that when A thinks about or experiences X, she thinks about or experiences Y and attributes some connection between X and Y.

So, for example, a drawing can be reminiscent of my nephew, the smell of honeysuckle can be reminiscent of golfing, etc.

There are three noteworthy features of reminiscence. First, the representational nature of reminiscence can be reduced to the representational nature of more fundamental entities. For example, I can explain the drawing's reminiscence of my nephew in virtue of the mental state produced by the drawing (which is about my nephew, who created it). Second, stipulation is sufficient to create an instance of reminiscence. For example, I could draw a symbol on my hand which I create for the sake of reminding me to buy bread from the store. The reminiscent relationship exists because of my stipulative act. Finally, any limitations of reminiscent relationships will be made for pragmatic reasons. For example, it would be for pragmatic reasons that I make the symbol on my hand look like a loaf of bread.

[2] I should note that the account of reminiscence here is not meant as a detailed explanation of this concept, but only as an analogy to draw a point about representation.

These three features of reminiscence are noteworthy because they are shared by Callender and Cohen's view of scientific representation. In fact, from Callender and Cohen's perspective, the only major difference between the two concepts would be the particular aims for which each relationship is utilized. While important, these different aims alone are insufficient to explain a key dissimilarity between scientific representation and reminiscence: while reminiscence can be private, scientific representation is necessarily communal. That reminiscence can be private can be seen from the fact that discussions of reminiscence can terminate in disagreement. For example, no one is ultimately 'correct' about whether or not someone is reminiscent of someone else. This is because reminiscence is agent-relative and so depends only upon some particular agent and her mental states.

Scientific representation relies on much more. As Suárez has argued, "representation is not at all 'in the mind' of any particular agent. It is rather 'in the world', and more particularly in the social world – as a prominent activity or set of activities carried out by those communities of inquirers involved in the practice of scientific modelling" (2010, 99). Scientific representation is not isolated from the practice in which it is embedded. It is necessarily communal.[3] The communal nature is demonstrated from the fact that representational vehicles demonstrate autonomy from individual scientists and their mental

[3] The view of representation argued for in this paper echoes many of the points made by Ludwig Wittgenstein's in his 'Private Language Argument' where he argues that meaning is necessarily communal (1953/2009, 95^{e}-111^{e}).

states.[4] For example, a scientist's rogue stipulation that the Lotka-Volterra model (which represents predator-prey relations) represents population change due to genetic drift does not count as an instance of scientific representation. This is not only because it does not (pragmatically) allow for meaningful insights, but also because it ignores and discounts the autonomous elements of the model as understood by the broader scientific community.[5] The autonomous elements are seen in the materiality or historicity of the representational vehicle; in its development, reception, and contemporary use. Understanding how and why the scientific object represents its target requires paying attention to these communal features. That is to say that the communal nature is partially *constitutive* of the representational relationship. Callender and Cohen's account of scientific representation does not sufficiently account for these constitutive communal elements, as will be shown more explicitly below.

4. Licensing

Explaining the communal nature of scientific representation requires that attention be given to the material, autonomous dimensions of the representational vehicle in terms of its

[4] This point has already been made specifically with regard to models by Morrison and Morgan (1999). Here, I am extending a similar point to other representational vehicles, including things like diagrams and figures.

[5] Of course, there may be disagreements and developments internal to the practice about how to use some representation, but these disagreements and developments are *part of the practice*.

development, reception, and use. All of these features partially establish a scientific representation, through an activity I call *licensing*. Licensing is the set of activities of scientific practice by which scientists establish the representational relationship between a vehicle and its target. It is itself a constitutive element of the representational relationship: it is a critical part in explaining how and why some vehicle represents its target. Seeing the sorts of activities involved in licensing and how they partially constitute the representational relationship will require that we pay close attention to the historical development, reception, and use of actual instances of scientific representation.

4.1 Licensing in Artistic Representation

A similar sort of licensing is present in representation in art, and so an initial pass on the concept as it applies to artistic practice will be helpful to draw an analogy to licensing in science.[6] To see the role of licensing in artistic representation, consider an example. The mere stipulation that Pablo Picasso's *Guernica* should represent the pain of cyberbullying is clearly insufficient to make it represent this target. Understanding how *Guernica* is representational involves an awareness of communal features: Picasso's intentions within the environment in which he created the painting, how the painting was received by viewers in the years following its creation, and how it is understood today. With these features in mind,

[6] It is somewhat contentious to draw conclusions about the nature of representation in science by appeal to art; see e.g. Bueno and French (2011). Nonetheless, it is a common technique in discussions of scientific representation; see e.g. Suárez (2004).

it is clear that *Guernica* represents the pain and suffering of the people of Guernica who had been bombed by axis forces at the request of Francisco Franco and the Spanish Nationalists. The licensing here is a constitutive element of *Guernica*'s representational nature: without these features, it is not clear whether or how the painting would manage to represent anything at all.

Licensing also occurs outside of the scope of authorial intent, when the artistic community comes to accept that a piece of art is representational in a way that was not intended by the author. A good example can be taken from an anecdote related by the author Flannery O'Connor:

> [A] student asked me…: "Miss O'Connor, what is the significance of the Misfit's hat?" Of course, I had no idea the Misfit's hat was significant, but finally I managed to say, "Its significance is to cover his head." (1988, 853)

The Misfit is a key character in O'Connor's famous short story, "A Good Man is Hard to Find," and, as such, it would not be surprising for his wardrobe to be importantly representational. Her answer indicates that while she did not intend any representational target for the hat, there may yet be one. If the hat is representational, it will not be due to her authorial intent, but rather due to the views of the broader artistic community.

Let me make it very clear that the licensure so far described is not already accounted for by elements of Callender and Cohen's account. First, notice that none of these means of licensing is a mere pragmatic limitation of already existing representations. It is not as if *Guernica* represents anything and everything, but is then *limited* by the contexts of Picasso,

audiences, and art historians. These contexts are a crucial part of understanding why it represents at all. Nor is the licensing mere stipulation. O'Connor leaves it open that there may be a representational target for the Misfit's hat, even though she did not stipulate one. A single reader's stipulation alone is insufficient to make it a representation, since the target must also fit well with the Misfit's characteristics, with O'Connor's general themes as understood by literary critics and audiences alike, and so on. Once again, these contexts are a critical part of establishing the representational nature of the hat.

4.2 Licensing in Scientific Representation: A Case Study

The unique aims of science indicate that the licensing of scientific representation is of a different kind than the licensing in art. All the same, licensing similarly plays a critical role in establishing scientific representation. According to Tarja Knuuttila, case studies of scientific representation have revealed that it is "a complicated phenomenon" and "a laborious art" (2014, 304). Understanding the nature of licensing and its role in the complexities of scientific representation will be best accomplished by examining the complicated features seen in the context of a case study. Examples could be made of any type of representational vehicle, like the masterful case study of a scientific figure made by Bruno Latour (1999). I will take as my example the Lotka-Volterra model, since its development exhibits interesting features, many of which have already been widely discussed by other philosophers (e.g. Knuuttila and Loettgers 2011, forthcoming).

As mentioned above, the Lotka-Volterra model is used by ecologists to represent predator-prey relations. It had its beginnings in the independent work of two different

scientists, Vito Volterra and Alfred Lotka. In understanding the representational nature of this model, it is important to pay attention to the licensing through its historical development. This attention includes noticing things like the way that the construction of the model by Lotka, Volterra, and others has been responsive to certain theoretical and empirical aims. These historical and practice-centered features of the model's development reveal the partial autonomy of its representational nature. These features constitute the licensing which is itself partially constitutive of the representational nature of the model since understanding how and why the model represents its targets requires attending to these features. Let us now turn to examine these features in more detail.

Consider first the development of the model by Volterra, who was "motivated by the goal of reproducing the kind of oscillating behavior that was observed empirically in fishery statistics" (Knuuttila and Loettgers forthcoming, 19). His aim to address a theoretical question with an empirically useful model is central not only to understanding how the model historically came about, but in understanding how it represents its targets. Consider how Volterra described his project and the aims which permeate his description:

> Let us seek to express in words the way the phenomenon proceeds roughly: afterwards let us translate these words into mathematical language. This leads to the formulation of differential equations. If then we allow ourselves to be guided by the methods of analysis we are led much farther than the language and ordinary reasoning would be able to carry us and can formulate precise mathematical laws. These do not contradict the results of observation. Rather

> the most important of these seems in perfect accord with the statistical results. (1928, 5)

Volterra's actual process of moving from words, to equation, to application of results (for both theoretical and empirical purposes) first involved creating an equation to account for the population change of a single species. He then added additional species and modelled interactions under different conditions, including, notably, contending for the same food and the predation of one species upon the other. Using these models, he demonstrated "three fundamental laws of the fluctuations of the two species living together" (1928, 20). He then applied these theoretical laws of predator-prey relations to the empirical case which had prompted his analysis, the peculiar rise in predator populations during the decrease of fishing of prey populations in the Adriatic Sea during World War I (1928, 21).

Why does Volterra's model represent these theoretical features of predator-prey relations? Why does it represent the populations of fish in the Adriatic during World War I? It represents these targets because, through a series of steps of analysis, revision, and development, each of which was responsive to certain theoretical and empirical aims understood and described in his account, Volterra *established* this representational nature. Indeed, as explained by Knuuttila and Loettgers (forthcoming), the historical development of this model has a much more extended history than the one Volterra described in the two papers where he first introduced it (1926, 1928). The model is a representation of its target not by mere stipulation and pragmatic constraint, but through careful and attentive construction of equations which ensure that the model functions in the wider theoretical

contexts and can explain the relevant empirical aims. In short, the model represents its targets because Volterra so *licensed* it by building into the model these external, autonomous representational features. Without these features, how or what would it represent?

Consider another instance of licensing in the development of the Lotka-Volterra model, this time by Lotka. His development proceeded with a different aim than Volterra: "instead of starting from the different simple cases and generalizing from them, he developed a highly abstract and general model template that could be applied in modelling various kinds of systems" (Knuuttila and Loettgers forthcoming, 13). He began by creating a very general equation which described "evolution as a process of redistribution of matter among the several components…of the system" (Knuuttila and Loettgers forthcoming, 15). In two papers (1920a, 1920b), Lotka applied this general equation to particular cases in biology and chemistry, in each case coming to theoretical conclusions about the systems in question. For example, in applying the equation to a predator-prey system, he concluded that there would be "undamped oscillation continuing indefinitely" among the two populations (1920a, 414). Lotka did not specifically apply the results to any empirical data, but instead used his results to come to theoretical conclusions about these relationships which he then connected to theoretical ecological principles drawn from Herbert Spencer's *First Principles* (1920a, 414).

Why does Lotka's model represent its theoretical target? What constitutes this representational relationship? Any attempt to explain the representational relationship must reference the way in which Lotka derived his general equation and the way in which he applies it to the specific cases. That is to say, the representational nature of the model is

constructed through the scientific activities performed by Lotka during the development of the model. Lotka does not merely stipulate that his model targets predator-prey relationships. Instead, he builds this ability into the model during the development of the general equation and further constructs this ability in his application of the question to specific targets. In so doing, he partially constructs the representational nature of the model—he licenses it as a representation through activities in accord with the broader practice.

The Lotka-Volterra model's history since its initial development is long and complex. As described by Alan Berryman (1992), one development was a shift in the 1940s to the use of a logistic formulation which allowed for attention to be placed on predator-prey ratios rather than products. Another development, which occurred around the same time, was the use of a predator functional response which introduced a nonlinear rate of death for the prey. These developments license new representational targets by expanding and altering the model to make it responsive to different theoretical or empirical aims, by removing idealizations, or otherwise by allowing for different theoretical conclusions. Many other variations of the Lotka-Volterra model exist, licensed by similar developments. Additionally, the original formulation of the model is still used in introductory textbooks on ecology (see, e.g. Cain, Bowman, and Hacker 2008). The representational nature of the model in each of these cases is partially established by these features of the model which stand independent of any mental states of scientists and students alike. In short, the constitution of the representational nature of the Lotka-Volterra model relies deeply upon these historical features of licensing as understood by the broader scientific community.

Let me briefly underscore the importance of these activities of licensing to the representational nature of the Lotka-Volterra model by imagining a scenario in which these features are absent. Suppose that Volterra and Lotka had proceeded differently. Suppose that they began, for no particular reason, by drawing a five-pointed star and stipulated that it represented predator-prey relations. What is the status of this star, qua representation? It is not as if the star *really* is a scientific representation of predator-prey relations albeit a bad representation (because it does a poor job of meeting certain pragmatic constraints). Rather, the star plainly fails to be a scientific representation at all. Scientific representations are constructed to assist in answering certain questions, explaining certain phenomena, understanding certain target systems. It is through licensing that scientists build into the vehicle the features capable of achieving these aims. A vehicle without licensing does not have this ability and so it is not just a bad representation. It is not a representation *at all*. Indeed, a discussion of the representational nature of vehicles which lack these features is either infelicitous or involves an equivocation of the word 'representation.' A view of scientific representation which equally counts both the star and the Lotka-Volterra model as full scientific representations, even if it specifies one as good and one as bad, underestimates the role of these historical features of the model. They are not external to the representational nature of the vehicle, but are themselves an essential constitutive feature of this representational nature: without these features, the vehicle is not a scientific representation at all.

5. The Special Problem of Scientific Representation

If I am right that licensing is a necessary constitutive feature of scientific representation which explains its communal nature, then contrary to Callender and Cohen's suggestion, we cannot pull the question of the constitution of representation away from questions of practice. A scientific object represents its target not (only) because there is some stipulation and pragmatic constraint, but also in virtue of licensing: the context in which it was created, the application of theoretical and empirical constraints, the awareness of and management of idealizations, and the history of its reception and use. Accounting for whether and how a scientific object represents its target will always require reference to these features which partially establish the representational nature. Thus, there *is* a special problem of scientific representation.

I should note that I am not here arguing for a stronger counter claim to Callender and Cohen which says that accounts of the representational nature of mental states are without *any* value to the constitution question of scientific representation. But my argument does indicate that an account of the representational nature of mental states *alone* is insufficient to account for scientific representation. Even if tomorrow we had a solid, universally accepted account of the representational nature of mental states, we would not yet have a complete account of scientific representation. We would still need an account of the deep reliance that it has upon the practice in which it is embedded. Thus, while our discussion of the constitution of scientific representation might include reference to the representational nature of mental states, it must also include reference to what I have described here as the licensing by the practice.

A different concern is that the use of the word 'special' is a bit deceptive. What I have identified here as the 'special' problem of scientific representation turns out to be a common feature of representation across disciplines, since, for example, I have suggested that it holds of artistic representation as well. While it is true that, according to my argument, an account of artistic representation will likely take account of licensing as well, it does not indicate that it is the *same type* of licensing in both practices. Indeed, given the unique aims that mark off scientific practice, its licensing can reasonably be expected to be correspondingly unique. That is to say that understanding, knowing, or explaining the empirical world are special aims, and therefore subject to special sorts of licensing. Scientific representation remains special because these features merit special attention.

We might also wonder whether it is right to continue to discuss scientific representation as a whole. If understanding representation in science requires in part that we understand the way in which scientists of a practice develop, utilize, and adapt these representational devices, then it is at least possible that these activities will be different within different domains. For example, the licensure of representations in physics might be rather different from that of economics. My suspicion is that, given the common broad scale aims of the various domains, we can still say some general things about representation in science as a whole. Nonetheless, we would do well to pay attention to representation as it occurs in these more localized contexts. Moving forward from this conclusion to develop further insights about the nature of scientific representation will involve analyzing specific representational objects or strategies as they occur in scientific practice, perhaps taking hints and clues from

in-the-field investigations like those conducted by sociologists of science, e.g. those in Lynch and Woolgar (1990), Latour (1999), and Coopmans et al. (2014).

6. Conclusion

Though Callender and Cohen's view remains a formidable approach to the constitution question of scientific representation, I have endeavored in this paper to show why their account is insufficient, and thus why this question merits continued attention by philosophers of science. Representation in science is deeply tied up with the practice in which it is embedded. The communal nature of scientific representation can be seen in the way that science, as a practice, partially constructs its representations through the activities of licensing. The licensing is not the pragmatic limitation of some already existing representations, but is itself a constitutive element of the representational relationship. Any account of what it is for a scientific object to represent its target will necessarily involve reference to licensing. Thus, there *is* a special problem of scientific representation.

Bibliography

Berryman, Alan. 1992. "The Origins and Evolution of Predator-Prey Theory." *Ecology* 73: 1530-1535.

Boesch, Brandon. 2015. "Scientific Representation." *Internet Encyclopedia of Philosophy*. http://www.iep.utm.edu/sci-repr/

Bueno, Otávio, and Steven French. 2011. "How Theories Represent." *British Journal for the Philosophy of Science* 62: 857-894

Cain, Michael, William Bowman, and Sally Hacker. 2008. *Ecology*. Sunderland, MA: Sinauer Associates, Inc.

Callender, Craig, and Jonathan Cohen. 2006. "There Is No Special Problem About Scientific Representation." *Theoria* 21: 67-85.

Contessa, Gabriele. 2007. "Scientific Representation, Interpretation, and Surrogative Reasoning." *Philosophy of Science* 74: 48-68.

Coopmans, Catelijne, Janet Vertesi, Michael E. Lynch, Steve Woolgar (eds.). 2014. *Representation in Scientific Practice Revisited*. Cambridge, MA: MIT Press.

French, Steven and James Ladyman. 1999. "Reinflating the Semantic Approach." *International Studies in the Philosophy of Science* 13: 103-119.

Giere, Ronald. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

———. 2004. "How Models Are Used to Represent Reality." *Philosophy of Science* 71: 742-752.

Hughes, R.I.G. 1997. "Models and Representation." *Philosophy of Science* 64 (Proceedings): S325-S336.

Knuuttila, Tarja, and Andrea Loettgers. 2011. "The Productive Tension: Mechanisms Vs. Templates in Modeling the Phenomenon." In *Models, Simulations, and Representations*, ed. P. Humphreys and C. Imbert, 3-24. New York: Routledge.

———. Forthcoming. "Modelling as Indirect Representation? The Lotka-Volterra Model Revisited." *British Journal of Philosophy of Science*, in press.

Knuuttila, Tarja. 2014. "Reflexivity, Representation, and the Possibility of Constructivist Realism." In *New Directions in the Philosophy of Science,* ed. M. C. Galavotti, S. Hartmann, M. Weber, W. Gonzalez, D. Dieks, and T. Uebel, 297-312. Dordrecht, The Netherlands: Springer.

Latour, Bruno. 1999. Circulating Reference. In *Pandora's Hope.* Cambridge: Harvard University Press.

Lotka, Alfred. 1920a. "Analytical Note on Certain Rhythmic Relations in Organic Systems." *Proceedings of the National Academy of Arts and Sciences* 42: 410-415.

———. 1920b. "Undamped Oscillations Derived from the Law of Mass Action." *Journal of the American Chemical Society* 42: 1595-1598.

Lynch, Michael E., and Steve Woolgar (eds.). 1990. *Representation in Scientific Practice*. Cambridge: MIT Press.

Morgan, Mary, and Margaret Morrison (eds.). 1999. *Models as Mediators: Perspectives on Natural and Social Science.* New York: Cambridge University Press.

O'Connor, Flannery. 1988. "The Catholic Novelist in the Protestant South." In *Flannery O'Connor: Collected Works,* 853-864. New York: Literary Classics of the United States.

Suárez, Mauricio. 2004. "An Inferential Conception of Scientific Representation." *Philosophy of Science* 71: 767-779.

———. 2010. "Scientific Representation." *Philosophy Compass* 5: 91-101.

———. 2015. "Representation in Science." In *The Oxford Handbook of Philosophy of Science*, ed. P. Humphreys. New York: Oxford.

van Fraassen, Bas C. 1980. *The Scientific Image*. New York: Oxford University Press.

———. 2008. *Scientific Representation: Paradoxes of Perspective*. New York: Oxford University Press.

Volterra, Vito. 1926. "Fluctuations in the Abundance of a Species Considered Mathematically." *Nature* 128: 558-560.

———. 1928. "Variations and Fluctuations of the Number of Individuals in Animal Species Living Together." *Journal du Counseil International Pour l'Exploration de la Mer* 3:3-51.

Wittgenstein, Ludwig. 1953/2009. *Philosophical Investigations,* trans. G.E.M. Anscombe, P. M. S. Hacker, and J. Schulte. Malden, MA: Wiley-Blackwell.

Dissolving the missing heritability problem

Abstract: Heritability estimates obtained in genome-wide association studies (GWAS) are much lower than those of traditional quantitative methods. This has been called the "missing heritability problem". By analyzing and comparing these two kinds of methods, we first show that the estimates obtained by traditional methods involve some terms that GWAS do not. Second, the estimates obtained by GWAS do not take into account epigenetic factors transmitted across generations, whilst they are included in the estimates of traditional quantitative methods. Once these two factors are taken into account, we show that the missing heritability problem can be largely dissolved. Finally, we briefly contextualize our analysis within a current discussion on how non-additive factors relate to the heritability estimates in GWAS.

1. Introduction.

One pervasive problem encountered when estimating the heritability of quantitative traits is that the estimates obtained from Genome-Wide Association Studies (GWAS) are much smaller than that calculated by traditional quantitative methods. This problem has been called the missing heritability problem (Turkheimer 2011). Take human height for example. Traditional quantitative methods deliver a heritability estimate of about 0.8, while the first estimates using GWAS were 0.05 (Maher 2008). More recent GWAS methods have revised this number and estimate the heritability of height to be at most 0.45 (Yang et al. 2010; Turkheimer 2011). Yet, half of the heritability is still missing.

In quantitative genetics, heritability is defined as the portion of phenotypic variation in a population that is caused by genetic difference (Downes 2015). Traditionally, this portion is estimated by measuring the phenotypic resemblance of genetically related individuals without identifying at the molecular level (more particularly the DNA level) the genetic causes of phenotypic variation. GWAS have been developed in order to locate the DNA sequences that influence the target trait and estimate their effects, especially for common complex diseases such as obesity, diabetes and heart disease (Visscher et al. 2012; Frazer et al. 2009). As for height, almost 300 000 common DNA variants in human populations that associate with it have been identified by GWAS (Yang et al. 2010). Granted by many that the heritability estimates obtained

by traditional quantitative methods are quite reliable, the method(s) used in GWAS have been questioned (Eichler et al. 2010).

A number of partial solutions to the missing heritability problem have been proposed, with most of them focusing on improving the methodological aspects of GWAS in order to provide a more accurate estimate (e.g., Manolio et al. 2009; Eichler et al. 2010). Some authors have also suggested that heritable epigenetic factors might account for part of the missing heritability. For instance, in Eichler et al. (2000, 488), Kong notes that "[e]pigenetic effects beyond imprinting that are sequence-independent and that might be environmentally induced but can be transmitted for one or more generations could contribute to missing heritability." Furrow et al. (2011) also claim that "[e]pigenetic variation, inherited both directly and through shared environmental effects, may make a key contribution to the missing heritability." Others have made the same point (e.g., McCarthy and Hirschhorn 2008; Johannes et al. 2008). Yet, in the face of this idea one might notice what appears to be a contradiction: how can *epigenetic* factors account for the missing heritability, if the heritability is about *genes*?

To answer this question as well as to analyze the missing heritability problem, we compare the assumptions underlying both heritability estimates in traditional quantitative methods and those in GWAS. We argue that a) the heritability estimates of traditional methods include some terms associated with broad-sense heritability (H^2), as opposed to narrow-sense heritability (h^2); b) although GWAS are supposed to get h^2, h^2 relies on an evolutionary concept of the gene

that can include epigenetic factors while heritability estimates obtained from GWAS do not. With these two points being illustrated, we expect the missing heritability problem to be largely dissolved as well as setting the stage for further discussions.

The reminder of the paper will be divided into three parts. First, we briefly introduce how heritability is estimated in two traditional methods, namely twin studies and parent-offspring regression. We show that the estimates obtained by each methods include *some* non-additive elements and consequently correspond neither to H^2 nor to h^2, but to a notion in between which we term "broader-sense heritability". Second, we outline the basic rationale underlying GWAS and illustrate that they estimate heritability by considering solely DNA variants. By arguing that the notion of additive genetic variance does not necessarily refer to DNA sequences but can also refer to epigenetic factors in traditional quantitative methods, we show that the notion of heritability estimated in GWAS is more restrictive than that of traditional quantitative methods, and term this notion "DNA-based narrow-sense heritability". Finally, in Section 4, based on the conclusions from Section 2 and Section 3, we claim that the gap between the heritability estimates of traditional quantitative methods and those of GWAS can be explained away in two major ways. One consists in recognizing that if non-additive variance was removed from the estimates obtained via traditional methods, they would be lower. The other consists in recognizing that if epigenetic factors were taken into account by GWAS, the heritability estimates obtained would be higher. We conclude Section 4 by showing how our analysis sheds

some light on a discussion about the role played by non-additive factors in the missing heritability problem. Because human height has been "the poster child" of the missing heritability problem (Turkheimer 2011, 232), we will use this example to illustrate each of our points.

2. Heritability in Traditional Quantitative Methods.

According to quantitative genetics, the phenotypic variance (V_P) of a population can be explained by two components, its genotypic variance (V_G) and its environmental variance (V_E). In the absence of gene-environment interaction and correlation, we thus have:

$$V_P = V_G + V_E \quad (1)$$

From there broad-sense heritability (H^2) is defined as:

$$H^2 = \frac{V_G}{V_P} \quad (2)$$

V_G can further be portioned into the additive genetic variance (V_A), the dominance genetic variance (V_D) and the epistasis genetic variance (V_I). We have:

$$V_P = V_A + V_D + V_I + V_E \quad (3)$$

where V_A is the variance due to hypothetical genes making an equal and additive contribution to the trait studied (e.g., height). V_D is the variance due to interactions between alleles at one locus for diploid organisms, and V_I is the variance due to interactions between alleles from different loci. V_D and V_I together represent the variance due to particular combinations of genes of an organism.

Since genotypes of sexual organisms recombine at each generation via reproduction, dominance and epistasis effects are not transmitted stably across generations, only additive genetic effects are. Therefore, V_A is the variance due to stably transmitted genetic effects. Narrow-sense heritability (h^2) measures to what extent variation in phenotypes is determined by the variation in genes transmitted from parent(s) to offspring (Falconer and Mackay 1996, 123). It is defined as:

$$h^2 = \frac{V_A}{V_P} \quad (4)$$

h^2 is important in breeding studies and is used by evolutionary theorists who are interested in making evolutionary projections of a trait within a population across generations.

To know h^2, both V_A and V_P must be known. V_P, for most quantitative traits (including height), can be directly estimated by measuring individuals. However, there is no direct way to estimate V_A in traditional quantitative methods. The traditional way to estimate it requires two elements. First, one needs a population-level measure of a phenotypic resemblance of family

relative pairs[1]. This measure is obtained by calculating the *covariance* of the phenotypic values for those pairs. The choice of what sort of relatives to use depends on what data is available. The second element is the genetic relation between family pairs. It indicates the percentage of genetic materials the pairs are expected to share. With these two elements, one can estimate how much the genes shared contribute to the phenotypic resemblance. In a large population with different phenotypes, one can then estimate how much the additive genetic difference contributes to phenotypic difference in this population, which estimates h^2.

For simplicity, traditional quantitative methods usually assume that there is neither gene-environment interaction nor correlation (Falconer and Mackay 1996, 131). Thus the covariance between the phenotypic values (e.g., height) of pairs equals to additive genetic covariance, dominant and epistasis genetic covariance, plus the environmental covariance. A general equation for traditional quantitative methods can be written as follows:

$$Cov(P_1, P_2) = Cov(A_1 + D_1 + I_1 + E_1, A_2 + D_2 + I_2 + E_2) =$$
$$Cov(A_1, A_2) + Cov(D_1, D_2) + Cov(I_1, I_2) + Cov(E_1, E_2) \qquad (5)$$

where indexes "1" and "2" represent the two family members for each pair studied. $Cov(P_1, P_2)$ is the covariance between the phenotypic values of one individual with the other.

[1] Or the mean values of their class (e.g., offspring) depending on the particular method used.

A, D, I and E represent additive effects, dominant effects, epistasis effects and environmental effects respectively.

The most commonly used traditional methods for estimating heritability are twin studies. In these studies one already knows that monozygotic twins share almost 100% of their genetic material while dizygotic twins about 50%. The environment is typically divided into the part of the environment that affects both twins in the same way (the shared environment, C) and the part of the environment that affects one twin but not the other (the unique environment, U) (Silventoinen et al. 2003). Hence, in the absence of interaction and correlation between C and U, we have:

$$E = C + U \quad (6)$$

Assuming epistasis effects to be negligible (a common assumption in twin studies), by inserting Equation (6) into Equation (5), we have:

$$Cov(P_{T1}, P_{T2}) = Cov(A_{T1} + D_{T1} + C_{T1} + U_{T1}, A_{T2} + D_{T2} + C_{T2} + U_{T2}) =$$
$$Cov(A_{T1}, A_{T2}) + Cov(D_{T1}, D_{T2}) + Cov(C_{T1}, C_{T2}) + Cov(U_{T1}, U_{T2}) \quad (7)$$

where indexes "T1" and "T2" represent the two twins for each twin pair studied. $Cov(P_{T1}, P_{T2})$ is the covariance between the phenotypic values of one twin with the other.

Because each twin's unique environment by definition is independent of that of the other twin, $Cov(U_{T1}, U_{T2})$ is zero for both monozygotic and dizygotic twins. Given that variance is a special case of covariance when the two variables are identical, and that for monozygotic twins A_{T1}, D_{T1}, and C_{T1}equal to A_{T2}, D_{T2}, and C_{T2} respectively, we can formulate the equation from Equation (7) as follows:

$$Cov_{MT}(P_{T1}, P_{T2}) = V_A + V_D + V_C \quad (8)$$

where $Cov_{MT}(P_{T1}, P_{T2})$ is the covariance between the phenotypic values of monozygotic twin pairs studied.

By contrast, dizygotic twins are expected to share half of their genes, which means that the covariance between the phenotypic values of one twin with the other of dizygotic twin pairs studied $(Cov_{DT}(P_{T1}, P_{T2}))$ is expected to be equal to half of the additive genetic variance, a quarter of dominant variance [2], and all of the shared environmental variance (with $Cov(U_{T1}, U_{T2})$ also to be zero). We have:

$$Cov_{DT}(P_{T1}, P_{T2}) = \frac{1}{2}V_A + \frac{1}{4}V_D + V_C \quad (9)$$

It is classically assumed that V_C in Equation (8) and (9) is the same. That is to say, for both monozygotic and dizygotic twin pairs, it is assumed that the shared environment would act in

[2] For each given gene with two alleles, the possibility that dizygotic twins have the same genotype is one quarter.

the same way if the pair has been reared together.[3] V_C can be cancelled by subtracting Equation (9) from Equation (8). The heritability can then be estimated as follows:

$$h^2_{b_{TS}} = \frac{2\{Cov_{MT}(P_{T1},P_{T2}) - Cov_{DT}(P_{T1},P_{T2})\}}{V_P} = \frac{V_A}{V_P} + \frac{\frac{3}{2}V_D}{V_P} \quad (10)$$

We call $h^2_{b_{TS}}$ broader-sense heritability (the index "b" is for "broader-sense") from *twin studies*, because the resulting estimate (which is about 0.8 for height) provides an accurate estimate of neither H^2 nor h^2, although it is closer to H^2 than to h^2 (Falconer and Mackay 1996, 172). That is to say, it corresponds to a definition of heritability that includes *some* elements of broad-sense heritability but not all of it.

Another often used traditional quantitative method to estimate heritability involves a parent-offspring regression. This method also assumes neither gene-environment interaction nor correlation, the covariance between the height of parents (one or the mean of both) and the mean of their offspring (Falconer and Mackay 1996, 164), equals to additive genetic covariance, dominant covariance (the epistasis covariance is assumed to be small and is not included), plus environmental covariance. Hence, Equation (5) can be formulated as follows:

[3] This assumption might be problematic because monozygotic twins are often treated more similarly by their parents than are dizygotic twins, and monozygotic twins are more likely to share a placenta than dizygotic twins. The difficulty can be mitigated by using adoption twin studies in which the environments for twins are random on average. But large adoption twins' data are exceedingly difficult to get (Griffiths 2005).

$$Cov(P_P, P_O) = Cov(A_P + D_P + I_P + E_P, A_O + D_O + I_O + E_O) =$$

$$Cov(A_P, A_O) + Cov(D_P, D_O) + Cov(E_P, E_O) \quad (11)$$

where indexes "P" and "O" represent the "parents" and the "offspring".

Two assumptions are then made. The first one is that there is no dominant effects transmitted from the parents to the offspring assuming the parents are unrelated (Doolittle 2012, 178), which means $Cov(D_P, D_O)$ is nil. Another assumption is that there is no correlation between the parents' environment and the offspring's environment so that $Cov(E_P, E_O)$ in Equation (11) is also nil. Given that on average, parents share in expectation 50% of genes with their offspring (parents and offspring share half of their genes), it leaves Equation (11) with a result of half of additive genetic variance ($\frac{1}{2}V_A$). Given V_P, h^2 can be estimated straightforwardly.

But the above two assumptions are problematic. First, the assumption of unrelated parents might be violated because of assortative mating in humans resulting in parents to be more genetically similar than two randomly chosen individuals (Guo et al. 2014). Hence, $Cov(D_P, D_O)$ is likely to be non-nil. Second, because the environments experienced by individuals are likely to be more similar within a family line, $Cov(E_P, E_O)$ might not be nil, either. If we take these two factors into consideration, the covariance of the parents and their

offspring is equal to half of additive genetic variance, *plus* a variance term representing effects due to dominance and similarities between environments. This can be written formally as:

$$Cov(P_P, P_O) = Cov(A_P, A_O) + Cov(D_P, D_O) + Cov(E_P, E_O) = \frac{1}{2}V_A + V_{D\&EC} \qquad (12)$$

where $V_{D\&EC}$ represents the variance due to some dominance and environmental correlation effects between the parents and the offspring studied.

The heritability can then be estimated by doubling the parent–offspring covariance in Equation (12) and dividing the total phenotypic variance of the population as follows:

$$h^2_{b_{POR}} = \frac{2Cov(P_P,P_O)}{V_P} = \frac{V_A}{V_P} + \frac{2V_{D\&EC}}{V_P} \qquad (13)$$

For similar reasons as with the heritability estimates from twin studies, we call $h^2_{b_{POR}}$ broader-sense heritability (with the index "b" also being for "broader-sense") from *parent-offspring regression*. Indeed, although it is often assumed that $h^2_{b_{POR}}$ represent h^2 (Falconer and Mackay 1996, 147), the resulting estimate (also about 0.8 for height) is broader than h^2 as it can include a component led by dominance variance and environmental correlation between parent and offspring.

To conclude this section, heritability estimates in both twin studies and parent-offspring regression include an extra term when compared to h^2, but they do not correspond to H^2. For this reason we regroup them under the term h^2_b for "broader-sense heritability", such that:

$$h_b^2 = h^2 + h_{other}^2 \quad (14)$$

where h_{other}^2 is the part of heritability contributed by the extra component(s) representing non-additive variance.

3. **Heritability in GWAS.**

Although any two unrelated individuals share about 99.5% of their DNA sequences, their genomes differ at specific nucleotide locations (Aguiar and Istrail 2013). Given two DNA fragments at the same locus of two individuals, if these fragments differ at a single nucleotide, they represent two variants of a Single Nucleotide Polymorphism (SNP). GWAS focus on SNPs across the whole genome that occur in the population with a probability larger than 1% which are called common SNPs. If one variant of a common SNP, compared to another one, is associated with a significant change on the trait studied, then this SNP is a marker for a DNA region (or a gene) that leads to phenotypic variation. For a polygenic trait like height, if we can detect all the SNPs that associate with it, then all the DNA difference makers that determine height difference can be located.

The development of commercial SNP chips makes it possible to rapidly detect common SNPs of DNA samples from all the participants involved in a study. By using a series of statistical tests, it can be investigated at the population level whether each SNP associates with

that target trait. The choice of the statistical tests depends on the data available as well as the trait studied. For quantitative traits like height, the most common approach is to make an analysis of variance table and assess whether the mean height of a group with one variant at one nucleotide is significantly different from the group with another variant of the same SNP[4] (Bush and Moore 2012). With all the SNPs associated with height being detected, data from the HapMap project, which provides a list of SNPs that are markers for most of the common DNA variants in human populations (Consortium, International HapMap 3 2010), is used to map the associated SNPs with common DNA variants. These mapped DNA variants, to be distinguished from DNA variants that do not affect the target trait, have been called "causal variants" (Visscher et al. 2012).

Based on the readings of SNP chips as well as further independent tests for SNPs, the effects of the associated SNPs (markers for causal DNA variants) on the trait can be calculated. By estimating the phenotypic variance contributed by these SNPs and the total phenotypic variance of the population, the heritability of causal DNA variants can be estimated as the ratio of the phenotypic variance caused by all the associated SNPs compared to the total phenotypic variance of the population (Weedon et al. 2008). Since it is common for biologists to assume

[4] For categorical (often binary disease/control) traits, the association test used involves measuring an odds ratio, namely the ratio of the odds of disease for individuals having a specific variant of a SNP, and the odds of disease for individuals who have another variant at the same locus. If the odds ratio of a common SNP is significantly different from 1, then that SNP is considered to be associated with the disease (Bush and Moore 2012).

that genes are only made up of pieces of DNA, it is thought that the variance obtained from all the causal DNA variants represent exactly the additive genetic variance, and the heritability estimated by GWAS should match narrow-sense heritability (h^2) (Yang et al. 2010; Visscher et al. 2006). However, the assumption that additive genetic effects are solely based on DNA sequences is problematic when faced with the evidence of epigenetic inheritance.

As was mentioned in Section 2, traditional quantitative methods for estimating heritability are based on measuring phenotypic values and genetic relations without reaching the molecular level. The genes are not defined physically, but functionally as heritable difference makers (Falconer and Mackay 1996, 123). In other words, they are theoretical units defined by their effects on the phenotype. With the discovery of DNA structure in 1953, it was thought that the originally theoretical genes were found in the physical DNA molecules. Since then, biologists commonly refer to genes as DNA molecules and this assumption is also made by researchers of GWAS. As [author] claim, this step was taken too hastily. If there is physical material, other than DNA pieces, that can affect the phenotype and be transmitted stably across generations, then it should also be thought to play the role that contributes to additive genetic effects.

Many studies have provided evidence for epigenetic inheritance[5], namely the stable transmission of epigenetic modifications across multiple generations and affect organism's traits

[5] We use the notion of "epigenetic inheritance" in the broad sense that refers to the inheritance of phenotypic features via causal pathways other than the inheritance of nuclear DNA (Griffiths and Stotz 2013, 112).

(e.g., Youngson and Whitelaw 2008; Dias and Ressler 2014). A classical example of this is the methylation pattern on the promoter of the agouti gene in mice (Morgan et al. 1999). It shows that mice with the same genotype but different methylation levels display a range of colors of their fur, and the patterns of DNA methylation can be inherited through generations causing heritable phenotypic variations. Epigenetic factors such as self-sustaining loops, chromatin modifications and three-dimensional structures in the cell can also be transmitted over multiple generations (Jablonka et al. 2014). Studies on various species suggest that epigenetic inheritance is likely to be 'ubiquitous' (Jablonka and Raz 2009).

The increasing evidence of epigenetic inheritance seriously challenges the restriction of the concept of the gene in the evolutionary sense to be materialized only in DNA. Relying on traditional quantitative methods, it is impossible to distinguish whether additive genetic variance is DNA based or based on other material(s). Some transmissible epigenetic factors, which are not DNA based, might *de facto* be included into the additive genetic variance used to estimate h^2. This extension of heritable units also echoes to the recent suggestion that genetic (assuming genes to be DNA based) and non-genetic heredity should be unified in an inclusive inheritance theory (Danchin 2013; Day and Bonduriansky 2010).

To apply the idea that some epigenetic factors can lead to additive genetic effects, the additive variance of them ($V_{A_{epi}}$) should be added to the additive variance of DNA sequences ($V_{A_{DNA}}$) to obtain V_A. Assuming there is no interaction between $V_{A_{epi}}$ and $V_{A_{DNA}}$, we have:

$$V_A = V_{A_{DNA}} + V_{A_{epi}} \quad (15)$$

Inserting Equation (15) to Equation (4) leads to:

$$h^2 = \frac{V_{A_{DNA}}}{V_P} + \frac{V_{A_{epi}}}{V_P} \quad (16)$$

Here we term the first term on the right side of Equation (16) "DNA-based narrow-sense heritability" (h^2_{DNA}), and the second term "epigenetic-based narrow-sense heritability" (h^2_{epi}), we thus have:

$$h^2_{DNA} = h^2 - h^2_{epi} \quad (17)$$

4. **Dissolving the Missing Heritability**.

As we mentioned it in Introduction, since the first successful GWAS was published in 2005 (Klein et al. 2005), there have been a lot of proposals for methodological improvements in GWAS (Manolio et al. 2009; Eichler et al. 2010). Studies have been conducted according to those proposals that permit to obtain higher heritability estimates. Examples include increasing the sample sizes which has resulted in more accurate estimates (e.g., Wood et al. 2014), considering all common SNPs simultaneously instead of one by one which has increased the heritability estimates of height from 0.05 to 0.45 (see Yang et al. 2010), and conducting meta-analyses which can lead to more accurate results when compared to single analysis (see Bush

and Moore 2012). Biologists have also suggested to search for SNPs with lower frequencies than 1% in order to account for a wider range of possible causal variants (Schork et al. 2009).

Aside from these partial improvements, our analysis reveals two reasons explaining away the missing heritability problem: a) In traditional quantitative methods, the heritability estimates include extra terms which are not presented in GWAS; b) In GWAS, heritability is estimated solely from causal DNA variants, while in traditional quantitative methods the additive effects contributed by epigenetic difference (h^2_{epi}) are *de facto* included in the estimates.

These two reasons can be shown formally. Using our terminology, missing heritability (MH) equals to the estimates obtained by traditional quantitative methods (h^2_b) minus the estimates obtained by GWAS (h^2_{DNA}), which are 0.8 and 0.45 respectively in the case of height. Thus we have:

$$MH = h^2_b - h^2_{DNA} \tag{18}$$

Replacing h^2_b and h^2_{DNA} by the right hand side of Equation (14) and (17), we obtain:

$$MH = h^2_b - h^2_{DNA} = h^2 + h^2_{other} - \left(h^2 - h^2_{epi}\right) = h^2_{other} + h^2_{epi} \tag{19}$$

Which means that the missing heritability results from the part of heritability originating from epigenetic factors stably transmitted across generations, plus the part of heritability originating from non-additives factors.

Our point that part of the missing heritability can be dissolved by considering non-additive effects echoes to the claim that almost all GWAS to date have focused on additive effects might be a reason for the missing heritability (McCarthy and Hirschhorn 2008). Although there is not enough data to confirm that non-additive effects do explain away some part of missing heritability, this claim appears again and again in discussions on the missing heritability problem (see for instance Maher 2008; Frazer et al. 2009; Gibson 2010; Kong 2010; Moore 2010). Yang et al. (2010, 565) disagree with this claim and respond that "[n]on-additive genetic effects do not contribute to the narrow-sense heritability, so explanations based on non-additive effects are not relevant to the problem of missing heritability."

We agree with Yang et al. (2010) that non-additive effects do not contribute to h^2. That said, because the heritability estimates obtained from traditional quantitative methods do not strictly correspond to h^2 but include some non-additive elements, non-additive effects cannot be dismissed as irrelevant for the missing heritability problem, though probably they are relevant in a way that both Yang et al. (2010) as well as their opponents did not consider.

5. **Conclusion**.

We have provided two ways in which the missing heritability problem can be explained away. First, heritability estimates from traditional quantitative methods (h_b^2) are overestimated when

compared to h^2. The resulting estimates would be smaller if the non-additive elements were eliminated. Second, heritability estimates from GWAS (h^2_{DNA}) are underestimated when compared to h^2 because they do not take into account the additive effects of epigenetic factors behaving like evolutionary genes. The resulting estimates would be larger if epigenetic factors were taken into account. We have voluntarily stayed away from the question of whether heritability should be defined strictly relative to DNA sequences or if it should encompass any factors behaving effectively like an evolutionary gene. Our inclination is that there is no principled reason to exclude non-DNA transmissible factors from heritability measures, but our analysis does not bear on this choice.

References:

Aguiar, Derek, and Sorin Istrail. 2013. "Haplotype Assembly in Polyploid Genomes and Identical by Descent Shared Tracts." *Bioinformatics* 29 (13): i352–i360.

Authors. Forthcoming. "The Evolutionary Gene and the Extended Evolutionary Synthesis." *British Journal for Philosophy of Science*.

Bush, William S., and Jason H. Moore. 2012. "Genome-Wide Association Studies." *PLoS Computational Biology* 8 (12): e1002822.

Consortium, International HapMap 3. 2010. "Integrating Common and Rare Genetic Variation in Diverse Human Populations." *Nature* 467 (7311): 52–58.

Danchin, Étienne. 2013. "Avatars of Information: Towards an Inclusive Evolutionary Synthesis." *Trends in Ecology & Evolution* 28 (6): 351–358.

Day, Troy, and Russell Bonduriansky. 2011. "A Unified Approach to the Evolutionary Consequences of Genetic and Nongenetic Inheritance." *The American Naturalist* 178 (2): E18–E36.

Dias, Brian G., and Kerry J. Ressler. 2014. "Parental Olfactory Experience Influences Behavior and Neural Structure in Subsequent Generations." *Nature Neuroscience* 17 (1): 89–96.

Doolittle, Donald P. 2012. *Population Genetics: Basic Principles*. Vol. 16. Springer Science & Business Media.

Downes, Stephen M. 2015. "Heritability." In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University.

Eichler, Evan E., Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore, and Joseph H. Nadeau. 2010. "Missing Heritability and Strategies for Finding the Underlying Causes of Complex Disease." *Nature Reviews Genetics* 11 (6): 446–450.

Falconer, Douglas S., and Trudy F. C. Mackay. 1996. *Introduction to Quantitative Genetics*. 4th edition. Longman: Benjamin Cummings.

Feil, Robert, and Mario F. Fraga. 2012. "Epigenetics and the Environment: Emerging Patterns and Implications." *Nature Reviews Genetics* 13 (2): 97–109.

Frazer, Kelly A., Sarah S. Murray, Nicholas J. Schork, and Eric J. Topol. 2009. "Human Genetic Variation and Its Contribution to Complex Traits." *Nature Reviews Genetics* 10 (4): 241–251.

Furrow, Robert E., Freddy B. Christiansen, and Marcus W. Feldman. 2011. "Environment-Sensitive Epigenetics and the Heritability of Complex Diseases." *Genetics* 189 (4): 1377–1387.

Griffiths, Anthony JF., Susan R. Wessler, Richard C. Lewontin, William M. Gelbart, David T. Suzuki, and Jeffrey H. Miller. 2005. *An Introduction to Genetic Analysis*. 8th edition. New York: W. H. Freeman.

Griffiths, Paul, and Karola Stotz. 2013. *Genetics and Philosophy: An Introduction*. Cambridge University Press.

Guo, Guang, Lin Wang, Hexuan Liu, and Thomas Randall. 2014. "Genomic Assortative Mating in Marriages in the United States." *PloS One* 9 (11): e112322.

Jablonka, Eva, Marion J Lamb, and Anna Zeligowski. 2014. *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Revised edition. MIT Press.

Jablonka, Eva, and Gal Raz. 2009. "Transgenerational Epigenetic Inheritance: Prevalence, Mechanisms, and Implications for the Study of Heredity and Evolution." *The Quarterly Review of Biology* 84 (2): 131–176.

Johannes, Frank, Vincent Colot, and Ritsert C. Jansen. 2008. "Epigenome Dynamics: A Quantitative Genetics Perspective." *Nature Reviews Genetics* 9 (11): 883–890.

Klein, Robert J., Caroline Zeiss, Emily Y. Chew, Jen-Yue Tsai, Richard S. Sackler, Chad Haynes, Alice K. Henning, John Paul SanGiovanni, Shrikant M. Mane, and Susan T. Mayne. 2005. "Complement Factor H Polymorphism in Age-Related Macular Degeneration." *Science* 308 (5720): 385–389.

Maher, Brendan. 2008. "Personal genomes: The Case of the Missing Heritability." *Nature News* 456 (7218):18–21.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, and Aravinda Chakravarti. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461 (7265): 747–753.

McCarthy, Mark I., and Joel N. Hirschhorn. 2008. "Genome-Wide Association Studies: Potential next Steps on a Genetic Journey." *Human Molecular Genetics* 17 (R2): R156–165.

Morgan, Hugh D., Heidi GE Sutherland, David IK Martin, and Emma Whitelaw. 1999. "Epigenetic Inheritance at the Agouti Locus in the Mouse." *Nature Genetics* 23 (3): 314–318.

Schork, Nicholas J., Sarah S. Murray, Kelly A. Frazer, and Eric J. Topol. 2009. "Common vs. Rare Allele Hypotheses for Complex Diseases." *Current Opinion in Genetics & Development* 19 (3): 212–219.

Silventoinen, Karri, Sampo Sammalisto, Markus Perola, Dorret I. Boomsma, Belinda K. Cornes, Chayna Davis, Leo Dunkel, Marlies De Lange, Jennifer R. Harris, and Jacob VB

Hjelmborg.2003. "Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries." *Twin Research* 6 (05): 399–408.

Turkheimer, Eric. 2011. "Still Missing." *Research in Human Development* 8 (3-4): 227-–241.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. "Five Years of GWAS Discovery." *The American Journal of Human Genetics* 90 (1): 7–24.

Visscher, Peter M., Sarah E. Medland, Manuel AR Ferreira, Katherine I. Morley, Gu Zhu, Belinda K. Cornes, Grant W. Montgomery, and Nicholas G. Martin. 2006. "Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings." *PLoS Genet* 2 (3): e41.

Visscher, Peter M., William G. Hill, and Naomi R. Wray. 2008. "Heritability in the Genomics Era—concepts and Misconceptions." *Nature Reviews Genetics* 9 (4): 255–266.

Weedon, Michael N., Hana Lango, Cecilia M. Lindgren, Chris Wallace, David M. Evans, Massimo Mangino, Rachel M. Freathy, John RB Perry, Suzanne Stevens, and Alistair S. Hall. 2008. "Genome-Wide Association Analysis Identifies 20 Loci that Influence Adult Height." *Nature Genetics* 40 (5): 575–583.

Wood, Andrew R, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, and Zoltán Kutalik. 2014. "Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height." *Nature genetics* 46 (11): 1173–1186.

Yang, Jian, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, et al. 2010. "Common SNPs Explain a Large Proportion of the Heritability for Human Height." *Nature Genetics* 42 (7): 565–569.

Youngson, Neil A., and Emma Whitelaw. 2008. "Transgenerational Epigenetic Effects." *Annual Review of Genomics and Human Genetics* 9: 233–257.

Scientific expertise, risk assessment, and majority voting

Thomas Boyer-Kassem*

Working paper: comments welcome, but please do not quote without permission.

February 29, 2016

Abstract

Scientists are often asked to advise political institutions on pressing risk-related questions, like climate change or the authorization of medical drugs. Given that deliberation will often not eliminate all disagreements between scientists, how should their risk assessments be aggregated? I argue that this problem is distinct from two familiar and well-studied problems in the literature: judgment aggregation and probability aggregation. I introduce a novel decision-theoretic model where risk assessments are compared with acceptability thresholds. Majority voting is then defended by means of robustness considerations.

Keywords: scientific expertise, risk, majority voting, robustness, decision theory

*TiLPS, Tilburg University, The Netherlands. Email: t.c.e.boyer-kassem@uvt.nl

1 Introduction

Scientists are often asked by political institutions to give expert advice on pressing questions. For instance, agencies that regulate medicines regularly resort to expert panels, and national scientific academies give advice to the government or to the assemblies. Even after discussing, scientific experts do not always agree on the answer, and when they do, they may disagree on the justification for this answer. How should decisions that involve risk assessments be taken and justified within scientific expert panels? This is the central question studied in this paper. As a matter of fact, many expert panels take decisions using the majority voting rule. This is for instance the case in advisory committees in the European and in the American agencies that grant medicines authorization, respectively the EMA and the FDA.[1] But is it the best decision rule? Is majority voting *on the final decision* the best way to aggregate different experts' opinions, and to track their reasons? This paper is restricted to cases in which the expert panel is asked to take a decision on only one binary question, for instance to answer the question "Is the risk-benefit ratio of some medicine worth it to be authorized for commercial use?". This simple case is already interesting as it corresponds to many real-life cases: some expert panels are constituted on the sole purpose of answering one specific question, or are asked to answer several but logically unrelated questions — e.g. decisions about different medicines.

To study this problem, I introduce a novel decision-theoretic model. The true/false decision is supposed to be taken by comparing a risk assessment a (typically, a probability) to a risk acceptability threshold t, e.g. "true" if and only if $a < t$. For simplicity, a and t are supposed to be in $[0, 1]$, but any quantity might go.[2] It is assumed that the n experts agree on the threshold value, but differ in their individual risk assessments a_k $(k = 1, \ldots, n)$ — or conversely, that they agree on the assessment, but disagree on the threshold value. Typically, the question asked to the expert panel is in the form "Is X's risk below t?". The problem studied in this paper is to determine how the individual a_k's should be aggregated in comparison with t, so as to give the group's answer to this question (I shall speak equivalently of the group's decision, or of the group's belief on whether the risk is below t). Compared to probability aggregation theory which studies the aggregation of probabilistic opinions, the novelty of this model lies (i) in the introduction of a threshold comparison which projects probabilities into a binary space, and (ii) in the fact that the group has to take a

[1] Cf. Hauray and Urfalino (2007), Urfalino and Costa (2015).

[2] Real quantities can be mapped to the interval $[0, 1]$, for instance with the function $x \to 1 - 1/(1 + x)$).

stand on one binary question only, and not on a more complex agenda. Compared to judgment aggregation theory which studies the aggregation of an interconnected set of beliefs, the novelty is that individuals do not just have true/false beliefs but probabilistic ones, even if the group is asked to express a true/false belief in the end. The present problem can be considered as a first bridge between these two existing frameworks. The best decision rule for our binary question is likely to depend on the details of characteristics of the question, of the experts, of the available knowledge, and on other details. My methodological approach is not to conduct a detailed study of particular cases, but to look at features which most (interesting) cases share, so as to find general properties of the best decision rule — what is meant by "best" shall be discussed too.

The main claims of this paper are the following. I argue that the framework of probability aggregation cannot help us solve the present problem (Section 2), because the aggregation problems it considers are too general. For the aggregation of scientific risk assessment on a specific question, a theory of its own is needed, and I try to sketch one here. I then argue that robustness considerations clearly legitimate majority voting on the final decision (Section 3). But when justifications for the decisions are sought, majority voting can lead to inconsistencies and the expert panel should aggregate on the reasons separately, before deriving logically its decision (Section 4). Overall, the case for the majority rule is thus a mixed one.

2 Probability Aggregation and Beyond

A standard requirement for a scientific expert panel is that it provides justifications for its decision. In the present model, the decision has to be consistent with the comparison between the risk assessment and the threshold, so a minimal justification is that the panel has a belief on the risk assessment (as all experts have a belief on the risk assessment, it would be weird that the panel claims to refuse the authorization while not being able to say that it believes that the risk assessment is above the threshold). So, our problem includes as a first step the aggregation of the individual risk assessments $\{a_k\}_{1\leq k\leq n}$ into a single group assessment a — deeper justifications for the group's decision are contemplated in Section 4. The group's decision is supposed to be consistent with this assessment, so pragmatically the easiest way to do so may be for the group to first aggregate the individual assessments, and then compare the result to the threshold.

Majority voting on the decision itself is a standard way for expert groups to take decisions, but it does not proceed in that way. Can it be objected that, within our model, it lacks the requirement that the group should be attributed a belief

on the risk assessment? No, for the following reason. The result of the majority vote is "true" if and only if a majority of agents vote "true", i.e. if and only if a majority of agents have a numerical assessment below the threshold, i.e. if and only if the median of the agents' assessments is below the threshold. In other words, the majority voting rule on the decision is equivalent to considering that the group's assessment is the median of the individual assessments. Hence, majority voting is in the race. What are the other challengers? A standard way to aggregate probabilities is to make averages. The linear average is defined as $\sum_k a_k$, and it can be generalized with weights $\omega_k \geq 0$ and $\sum_k \omega_k = 1$, as $\sum_k \omega_k a_k$, to take into account unequal degrees of expertise on the question.[3] Other averages are the geometric average or the harmonic average. Our problem is to determine which probability aggregation rule, followed by the threshold, is the best one in our problem. It is easy to see that these various probability aggregation rules can give different binary decisions for the group.[4]

Pooling probability functions has been studied for several years in the theory of probability aggregation (for surveys, cf. Dietrich and List forth., Martini and Sprenger forth., section 3). Can its results be used to select the best aggregation rule in our problem? I shall argue that unfortunately no. The framework of probability aggregation adopts an axiomatic method: it starts by stating several axioms which appear as desirable properties for the pooling function and then studies which function or aggregation rule, if any, satisfies them. The axioms considered in Dietrich and List's survey can be expressed in our case as:

- **Independence**: the group's probability a only depends on the individual probabilities a_k.
- **Unanimity preservation**: if all agents' probabilities a_k are the same, then the group's probability a is this one too.
- Three **Bayesian axioms**: if some information is learned by all individuals, then the group's decision changes by conditionalization on that event.

[3] It is akin to the iterated Lehrer-Wagner model which, starting from respect weights agent have to one another, provides a single probability for the group. However, the iterated Lehrer-Wagner model, and even more its normative interpretation, have been subjected to many criticisms (for a survey, cf. e.g. Martini and Sprenger forth. section 4). As a descriptive model, it is not useful for the present discussion.

[4] Consider for instance the median and the linear average, with three experts with $a_1 = a_2 = 0.04$, $a_3 = 0.10$, and $t = 0.05$. A majority voting on the decision gives a "true" as two experts on three assess the risk to be below the threshold. The linear average (with equal weights) is 0.06, which is higher than t, so this gives a "false".

The Independence axiom is automatically satisfied here, because our problem contains only one true/false answer, and there is no other probability on which a could depend. The three Bayesian axioms make sense in cases where the expert panel learns new information. In our problem, however, an extensive discussion has already taken place so no agent learns new information anymore, and the expert panel is not making any new inquiry. So the Bayesian axioms are not relevant in our case, and only the Unanimity preservation axiom expresses a desirable property for the aggregation rule.

An essential point to note is that a very large number of aggregation rules satisfy this axiom: the median, linear averaging, geometric averaging, and so on — actually, any convex function of the a_k. This illustrates the fact that a classical uniqueness result from the probability aggregation literature does not hold anymore: the well-known theorem by McConway 1981 and Wagner 1982, which states that linear averaging functions are the *only* independent and unanimity-preserving functions. The reason is that the theorem requires a set of at least three events, whereas our problem only considers two — e.g. the product is risky, with probability a_k, and the product is not risk, with probability $1 - a_k$. Considering a simpler agenda has widened the set of suitable aggregation rules, and no theoretical result from the literature can be used to pick the best one. More generally, the uniqueness and impossibility results from the theory of probability aggregation are useless for our problem. So, how scientific expert panels should aggregate risk assessments is not a simple problem that can be solved straightforwardly with the existing literature, which has focused on general problems with complex agendas, and has thus neglected more specific yet important questions. In the next section, I discuss other desiderata or axioms that we would like to impose on the aggregation rule.

3 Robustness Matters

Scientific risk assessment is supposed to meet some standards of reliability and objectivity, and the aggregation of these assessments should follow alike standards. In this spirit, I now introduce several new requirements for our aggregation rule. The aggregation rule should be sensitive to the right features of our problem, and not to the parasitic ones. It should favor objective features at the detriment of idiosyncrasies or unwanted values (for an analysis of the concept of objectivity, cf. Douglas 2004 — I refer to some of her distinctions below). In other words, the aggregation rule should be robust to some changes that we regard as irrelevant. In this section, I defend three dimensions of robustness that should be taken into account: the risk metrics, the level of detail, and the presence of strategical agents.

Several probability aggregation rules can be considered: linear averaging, geometric averaging, harmonic averaging, among others. As the forthcoming robustness discussion is similar for all the various averagings, I shall simplify it and consider only linear averaging, which shall be contrasted with the median. $\mathcal{R}_a$ denotes the aggregation rule that compares the threshold with the linear average (which thus stands for other averages), and $\mathcal{R}_m$ the aggregation rule that compares the threshold with the median of the individual assessments (which is equivalent to a majority vote on the decision itself).

3.1 Metrics

The formal model I have introduced relies on a quantitative scale — a and t are given numerical values in $[0, 1]$. How is this scale defined in real cases? My talking about probabilities has been only a matter of simplicity given the reduction of the problem to the $[0, 1]$ interval, and typical cases do not bear on well-defined probabilities or explicit scales. For instance, a standard question posed at an FDA advisory committee is "Does the overall risk versus benefit profile for X support marketing in the US ?"[5]. This question supposes that experts identify the risk versus benefit profile, and determine the value of the threshold under which a marketing is warranted. This can be done in a number of ways, and these are essentially value-laden questions[6] — what is acceptable or not has to do with extra-scientific values, and may also reflect the fact that an expert is risk-averse or risk-seeking. Overall, it makes sense to suppose that both the metrics scale and the threshold depend on the experts. Conversely, as the aggregation procedure is supposed to take place when the experts have extensively discussed, one can make the simplifying assumption that the same facts are known to all, and thus that the risk assessment is the same for all. In that way, our model actually applies in the setting in which a is common to all experts, but each has her own threshold t_k. The fact that the quantitative risk scale is not uniquely defined can be approached from a mathematical viewpoint: any scale can be reparametrized by applying any continuous bijection from $[0, 1]$ to $[0, 1]$, such as $x \mapsto x^2$.

These points make a hard time for the rule $\mathcal{R}_a$ (and other non linear averagings). First, from a practical viewpoint, the dependence of the risk scale metrics on the expert prevents the use of rules which take as inputs the numerical values of the risk assessments or of the threshold. For instance, is it even possible for a chairman to ask her colleagues "Please tell me your overall risk versus benefit acceptability threshold"

[5] Cf. Urfalino and Costa (2015, p.183).

[6] On the role of values in science more generally, and a critic of the value-free ideal, cf. Douglas 2009.

(or assessment), given that each expert may have her own scale? The rule $\mathcal{R}_m$, as it is equivalent to majority voting, needs not rely on input individual numerical values, and is thus safe from this criticism. Second, even if these practical difficulties could be overcome, some theoretical difficulties remain. Suppose a common scale has been adopted so that all experts can express their t_k. An aggregation rule that depends on the metrics of that common scale can give different outcomes according to the scale employed, as shown in Table 1. This dependence is a problem: which common scale should be chosen? (This is another aggregation problem!) Note that a variant of this problem exists even with a well-defined probability scale. For instance, let A be the event that a certain risk (e.g. carcinogenic substances in food) is responsible for more than 10 cases of cancer in 100,000 people during 1 year. The experts estimate the probability of A, $p(A)$. Consider now A' the event that the risk is responsible for more than 10 cases of cancer in 100,000 people during 10 years. Call $p(A')$ its probability. If the cancer cases are independent along the years, then $p(A') = 1 - (1 - p(A))^{10}$. Because the relation between $p(A)$ and $p(A')$ is not linear, taking the linear average of the experts assessments on A, and transforming it into an assessment on A', or taking the linear average of the experts assessments on A', does not give the same result. Which event A or A' is the more "natural" is not clear, and so much more for the right risk group assessment.

This gives good reasons to consider the following requirement: the aggregation rule should be insensitive to the metrics used to describe the problem, i.e. the assessment and the threshold. What should matter is just the relative position of the a and t_k, not their distance which can be due to some idiosyncratic value-laden judgments. This is requiring that the aggregation rule is more objective, under the sense of value-neutral objectivity as characterized by Douglas (2004, p. 460), which does not mean "free from all value influence" (as judging whether a risk benefit ratio is lower enough is bound to involve a value judgment), but takes a position "that is balanced or neutral with respect to a spectrum of values" (here, the balance is reached by taking into account only relative positions). The metrics robustness excludes the rule $\mathcal{R}_a$ which employs a linear average — Table 1 has shown a counter-

	t_1	t_2	t_3	a	Average t	$\mathcal{R}_a$	$\mathcal{R}_m$
x scale	.01	.01	.1	.05	.04	**False**	False
x^2 scale	0.0001	0.0001	.01	0.0025	0.0034	**True**	False

Table 1: Example in which the rule $\mathcal{R}_a$ gives different answers depending on the scale. The three experts have different thresholds t_k and a common risk assessment a.

example — but not $\mathcal{R}_m$ which relies on the median.[7]

3.2 Level of detail

Another argument for an aggregation rule that does not rely on a specific metrics comes from considerations of the level of detail in which the problem is described. So far, a continuous scale has been assumed, with numerical assessments in $[0, 1]$. Numerical discrete scales could also be used or even qualitative assessments only — it corresponds to decisions under uncertainty and not under risk. Consider for instance the case of the well-known IPCC Assessment Reports, that formulate a synthesis of existing scientific knowledge on climate change issues. The reports use a standardized vocabulary to express uncertainties, with several scales: some are qualitative (e.g. low/medium/high), others are quantitative (and use probabilities).[8] The historical trend has been to use more quantitative scales and less qualitative scales, but the latter have the advantage of being easily understandable by non-technical audiences, and thus should continue to be used in the future. Some qualitative and quantitative scales are in an explicit correspondence, as illustrated on Table 2. Writing an IPCC report involves synthesizing large amounts of scientific literature, so co-authors of a chapter may have different beliefs on the uncertainties associated with a finding. Whether they express their beliefs on a qualitative or on a quantitative scale, the way their beliefs are aggregated should be smooth and not vary abruptly (some very precise yet qualitative scales are conceivable), all the more than some explicit correspondence exist (Table 2). This is also a question of historically

[7]The comparability of scales is also discussed in Risse's (2004) political philosophy work, who also takes it as an argument for majority voting.

[8]Cf. e.g. the last report of the Working Group I, Stocker et al (2013, p. 138-142).

Term	Likelihood of the Outcome
Virtually certain	99–100 % probability
Very likely	90–100% probability
Likely	66–100% probability
About as likely as not	33–66% probability
Unlikely	0–33% probability
Very unlikely	0–10% probability
Exceptionally unlikely	0–1% probability

Table 2: Likelihood terms associated with outcomes used in the Fifth Assessment Report of the IPCC (Stocker et al 2013, p. 142).

consistency when switching from qualitative to quantitative scales.[9] Thus, a sound requirement is that the aggregation rule extends to formulations with discrete and qualitative scales. As the average of non-numerical and qualitative values is not defined, $\mathcal{R}_a$ does not satisfy this requirement. The median is defined on any kind of scale, and $\mathcal{R}_m$ satisfies the requirement. So only $\mathcal{R}_m$ is robust for the level of detail.

3.3 Bias and strategical votes

Not all experts are moved by epistemic goals only, and conflicts of interests can arise. For instance, numerous controversies have surrounded the FDA advisory committees along the years (Urfalino and Costa 2015, p. 168-169.) If a better selection of experts may be the solution, the decision rule used in the expert panel can also reduce the impact of bias agents.[10] With $\mathcal{R}_a$, an expert can strategically express a much lower risk of a medicine to influence the group's average — with a threshold at 10 %, she might express 0.1% instead of just 9%. The aggregation rule should be insensitive to such a strategical vote manipulation, and this is all the more important as the biased agent may have already influenced other agents during the preceding discussion. $\mathcal{R}_m$ is clearly robust in this sense, as an agent has the same influence whether her probability is just below the threshold or close to 0. This is not so for $\mathcal{R}_a$. This robustness requirement also makes the aggregation rule more objective, in the sense of detached objectivity (Douglas 2004, p. 459): one's personal values (allegiance to a firm) should not prevail on evidence (e.g. that the probability is 9%, as above).

Overall, the three robustness requirements considered here clearly favor $\mathcal{R}_m$ over $\mathcal{R}_a$. This provides a substantial justification for the traditional democratic rule in expert panels confronted with a binary decision. This result is a real departure from probability aggregation theory, in which linear averaging is justified on solid grounds. Narrowing the agenda and introducing a threshold has changed the solution to the aggregation problem.

[9] One may object that in the IPCC case the co-authors aggregate beliefs without a threshold comparison for a binary decision. Actually, thresholds are implicit: a finding which confidence is too low may not be mentioned. Anyway, the IPCC example can be seen as a mere illustration of the level of detail problem.

[10] Biased and extremist agents have been much studied in the literature of opinion dynamics (cf. for instance in Lorenz's 2007 survey), but not so in the literature of opinion aggregation.

4 Reasons

So far, a simplified model of scientific expert panels has been considered, one in which the group is asked to give a binary decision. As argued, the first step in justifying that decision consists for the panel to have a belief on the risk assessment, which is given by the median of the individual assessments in the case of $\mathcal{R}_m$. However, expert panels are often asked to provide a deeper justification. The question then arises of how the panel should aggregate its members views on this justification. In this section, I propose a novel but simple model for individual numerical assessment justification, in line with my previous threshold model.

Perhaps the most typical interpretation of the risk assessment a is that of a (subjective) probability. Suppose this probability is determined by m independent factors ($m \geq 2$). For instance, the risk associated with a medicine comes from m unrelated secondary effects. Then a is the probability that at least one risk factor triggers:

$$a = 1 - \prod_{j=1}^{m} (1 - a_j). \tag{1}$$

Each expert k is supposed to have her own assessment of each factor $a_{k,j}$ ($j = 1, \ldots, m$). Our problem is then to aggregate the $n \times m$ matrix of probabilities $a_{k,j}$, and to compare that result with the threshold.

As the m factors are independent, a sound requirement is to aggregate the individual assessments on them separately. How should that be done? Adapting the arguments from the previous section, one is lead to the conclusion that the panel should take the *median* of the individual assessments for each factor. However, there is a fundamental limitation to this, due to the previously mentioned theorem by McConway and Wagner's (cf. Section 2). Here is why. Requiring as above that the aggregation proceeds on each factor independently is just requiring the classical independence axiom. Another legitimate requirement is the classical axiom of unanimity preservation: if all experts agree on the risk assessment for one factor, then the panel should take this value as its own. As $m \geq 2$, all the conditions of the theorem by McConway and Wagner are fulfilled[11], so its conclusion apply: the only probability aggregation rule on the set of factors and on the overall decision is linear averaging. This reveals that, if groups use the median to determine both the independence factors' values and the overall risk (according to the above results), then it does not give a probability function and inconsistencies can arise. Table 3

[11]Each of the $m \geq 2$ factors can be triggered or not, so there are at least 4 events, which is higher than the 3 required in the theorem.

gives such an example. In other words, asking the expert panel to take stands on the reasons for its majority decision can lead it to change its decision.

Does it mean that our robustness defense of the median should be discarded? Not necessarily. The theorem by McConway and Wagner assumes that the experts aggregate their views *both* on the independent factors and on the overall risk assessment. But one can have the experts aggregate their views on the independent factors only. The overall risk assessment is then computed according to Equation 1, and the final decision is logically obtained from a comparison between this value and the threshold. In that way, experts do not vote on the final decision directly. This decision rule is a so-called premise-based rule.[12] Then, the linearity result of McConway and Wagner does not apply any more. The robustness considerations from the previous section do apply at the level of independent factors, and they recommend that the group takes the median of the individual assessments.

The present model of factors has assumed that there exists some common numerical scale, so that taking the median of individual assessments makes sense. However, the previous section has in part argued that such a scale may not always exist. In these cases, the present model of independent factors cannot apply. The theory of judgment aggregation offers a general framework for the aggregation of non-numerical reasons or justifications, with true/false beliefs (for reviews, cf. List 2012, Martini and Sprenger forth.). Applying in detail this framework to our problem of scientific justification would require another paper. A general result from this literature, however, is the discursive dilemma: majority voting on a set of true/false beliefs related in a logical way (here: reasons for the decision) may generate inconsistent collective judgments. This echoes our own finding about the median, which corresponds to majority voting in case of a threshold comparison. So whatever

[12]On this strategy more generally, see Cooke (1991), Bovens and Rabinowicz (2006), Hartmann and Sprenger (2012). Another solution to our problem could be the conclusion-based rule, i.e. aggregate only the views on the conclusion, but this is just like the previous section that we are trying to surpass.

Risk aspect	a_1	a_2	$a = 1 - (1 - a_1) \cdot (1 - a_2)$
Agent #1	0.01	0.01	0.0199
Agent #2	0.02	0.01	0.0298
Agent #3	0.01	0.02	0.0298
Median	0.01	0.01	0.0199 or 0.0298 ?

Table 3: A case in which the rule of the median can lead to inconsistencies. With a threshold at e.g. 0.025, the group's decision could be either true or false.

the scale, majority voting on all parts of the question is in great difficulty, and a premise-based solution should be adopted.

5 Conclusion

This paper has investigated the rationale for the majority rule that is often used in scientific expert panels, when dissent persists after discussion, and has looked for the best decision rule in this context. To this end, I have introduced a threshold probability model for individual decisions. Three main points have been shown in the paper: (1) the standard framework of probability aggregation is unable to solve our problem of risk aggregation. (2) robustness considerations clearly favor majority voting on the decision, i.e. comparing the threshold to the median of the individual risk assessments. (The robustness axioms I have advocated, which have been designed from considerations on scientific expert panel, could in return inspire social choice theory). (3) when a justification of the panel's decision is looked for, the median rule (corresponding to majority voting) can lead to inconsistencies. The promising route is to have the group aggregate on the reasons level, not on the final decision one. This should encourage scientific expert panels to divide questions from a logical viewpoints, and to take decisions on sub-problems instead of voting on the final decision directly. Current practices in advisory committees of the FDA and of the EMA could evolve in this respect. However, these claims have only been shown in quite simple and idealized models of decision-making. Future work is needed to investigate other models. These preliminary results have nonetheless cast some serious doubts on the majority voting rule only applied on the final decision.

Note finally the generality of the proposed model, which goes well beyond scientific expertise: the a and t variables can be interpreted as degrees of beliefs or as utility measures, within an epistemology or an economy framework.

References

Bovens, Luc and Wlodek Rabinowicz. 2006. "Democratic answers to complex questions. An epistemic perspective". *Synthese* 150: 131-153.

Cooke, Roger M. 1991. *Experts in Uncertainty. Opinion and Subjective Probability in Science*. Oxford University Press.

Dietrich, Franz and Christian List. Forthcoming. "Probabilistic Opinion Pooling". In *Oxford Handbook of Probability and Philosophy*, Oxford University Press.

Douglas, Heather E. 2004. "The Irreducible Complexity of Objectivity." *Synthese* 138:453-473.

Douglas, Heather E. 2009. *Science Policy and the Value-Free Ideal*. University of Pittsburgh Press.

Hartmann, Stephan and Jan Sprenger. 2012. "Judgment aggregation and the problem of tracking the truth." *Synthese* 187:209-221.

Hauray, Boris and Philippe Urfalino. 2007. "Expertise scientifique et intérêts nationaux. L'évaluation européenne des médicaments 1965-2000". *Annales HSS* 2: 273-298.

List, Christian. 2012. "The theory of judgment aggregation: an introductory review." *Synthese* 187:179-207.

Lorenz, Jan. 2007. "Continuous Opinion Dynamics under Bounded Confidence: A Survey." *International Journal of Modern Physics C* 18, 1819.

Martini, Carlo and Jan Sprenger. Forthcoming. "Opinion aggregation and individual expertise." In *Scientific collaboration and collective knowledge*, ed. by T. Boyer-Kassem, C. Mayo-Wilson and M. Weisberg, Oxford University Press.

McConway, Kevin J. 1981. "Marginalization and Linear Opinion Pools." *Journal of the American Statistical Association* 76(374): 410-414.

Risse, Mathias. 2004. "Arguing for Majority Rule". *The Journal of Political Philosophy* 12(1): 41-64.

Stocker Thomas .F. et al. 2013. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Urfalino, Philippe and Pascaline Costa. 2015. "Secret-Public Voting in FDA Advisory Committees." In *Secrecy and Publicity in Votes and Debates*, ed. Jon Elster, 165-194. Cambridge University Press.

Wagner, Carl. 1982. "Allocation, Lehrer models, and the consensus of probabilities." *Theory and Decision* 14: 207-220.

Responsiveness and robustness in the David Lewis signalling game

Carl Brusse and Justin Bruner

October 28, 2016

Abstract

We consider modifications to the standard David Lewis signalling game and relax a number of unrealistic implicit assumptions that are often built into the framework. In particular, we explore realistic asymmetries that exist between the sender and receiver roles. We find that endowing receivers with a more realistic set of responses significantly decreases the likelihood of signalling, while allowing for unequal selection pressure often has the opposite effect. We argue that the results of this paper can also help make sense of a well-known evolutionary puzzle regarding the absence of an evolutionary arms race between sender and receiver in conflict of interest signalling games.

1 Signalling games and evolution

Common interest signalling games were introduced by David Lewis (Lewis, 1969) as part of a game theoretic framework which identified communicative conventions as the expected solutions to coordination problems. In recent years, this has informed a growing body of work on the evolution of communication, incorporating signalling games into an evolutionary game theoretic approach to modelling the evolution of communication and cooperation in humans (Skyrms, 2010; Skyrms, 1996).

As the basis for game theoretic modelling of such phenomena, David Lewis signalling games are attractive in their intuitive simplicity and clear outcomes. They are coordination games of common interest between world-observing senders and action-making receivers using costless signals; in contrast to games where interests may differ and where costly signals are typically invoked. In the standard two-player, two-state, two-option David Lewis signalling game (hereafter the '2x2x2 game'), the first agent (signaller) observes that the world is in one of two possible states (state1 or state2) and broadcasts one of two possible signals (signal1 or signal2) which are observed by the second agent (receiver) who performs one of two possible actions (act1 or act2). If the acts match the state of the world (i.e. act1 if state1 or act2 if state2) then the players receive a greater payoff than otherwise.

Most importantly, though, the game theoretic results are unequivocal. There exist two Nash equilibria that are, in Lewis's words, signalling systems where senders condition otherwise arbitrary signalling behaviour on the state of the world, and receivers act on those signals to secure the mutual payoff. The two

systems only differ on which signal gets to be associated with each state of the world[1]. Huttegger (2007) and Pawlowitsch (2008) have shown that under certain conditions a signalling systems is guaranteed to emerge under the replicator dynamics, a standard model of evolution to be discussed further in section 4.

Of course the degree to which Lewis' approach makes sense is the degree to which we have confidence in the interpretation and application of such a highly idealised model to the more complex target systems. The obvious worry is that by introducing more realistic features into the model one might break or significantly dilute previous findings on the evolution of signalling.

Not surprisingly, then, recent work on Lewis signalling games has investigated the many ways in which such de-idealizations could occur. Some deviations from the standard Lewis signalling game include: more and varied states of the world, the possibility of observational error or signal error, noisy signals, partial deviation in interest between senders and receivers, the reception of more than one signal, and so on. Many such concerns are dealt with favourably in Skyrms (2010), and in work by others. For example Bruner et al. (2014) generalizes beyond the 2x2x2 case and Godfrey-Smith and Martinez (2013) and Godfrey-Smith (2015) mix signalling games of common interest and conflict of interest. One complication of the Lewis signalling game (particularly important for our purposes) is that signalling systems are not guaranteed in the simple 2x2x2 case when the world is biased. In other words, when the probabilities of the world being in state1 or state2 are not equal, a pooling equilibrium in which no communication occurs between sender and receiver is evolutionarily possible.

2 Symmetry breaking

The focus here will be with the idealisation that sender and receiver are equally responsive in strategic settings. Senders and receivers (in the evolutionary treatment of such games) are two populations of highly abstract and constrained agency roles: all that signallers do on observing the state of the world is send a signal, and the receivers must act as though the world is in one or other of the sender-observable states. Of those two roles, it is the restriction on receivers which is the more problematic.

Imagine for example a forager sighting a prey animal at a location inaccessible to her, but close enough to be acquired by an allied conspecific (who cannot observe the animal). In this case, it is easy for the first forager to slip into the signalling role and execute it, whistling or gesturing to her counterpart. To play the receiver role, however, the second forager has to actually re-orient their attention (to some degree) and attempt to engage in appropriate behaviour for the world-state the first has observed (e.g. prey is to the east or to the west, etc.).

The Lewis signalling model by design is constrained such that the receiver's actions are limited to just those acts associated with the sender's observed world-states. It is of course sensible to begin inquiry with as simple of a model as possible and consider a limited range of responses to stimuli. However, our point is that it is more plausible to make these idealizations for signallers than

[1]The other two possible outcomes of the game are 'pooling equilibrium', where the receiver plays act1 or act2 unconditionally.

for receivers. Signals are (by stipulation) cheap and easy to send, yet the actions available to the receiver are less plausibly interpreted as intrinsically cheap and free of opportunity cost.

In addition, the informational states drawn on by sender and receiver are also likely to be very different. Any real-life sender's observation of a world state will likely inform their motivations ('we should catch that animal') to dictate a fairly clear course of action ('try to direct the other agent's behaviour'). But all the receiver gets is a whistle, gesture or other signal which (by stipulation) has no pre-established meaning. The experience of observing a strategically relevant state of the world will typically be richer and more detailed than that of observing a strategically relevant artificial signal. All this leads to two concerns. Firstly, asymmetries in the strategic situations are likely to exist between senders and receivers. Receivers are likely to have locally reasonable options available to them other than those relevant to signaller-observed states of the world, and their responsiveness to the strategic situation is therefore less satisfactorily modelled by the strictly symmetric payoff structures of standard signalling games. Call this the structural responsiveness concern.

Secondly, given the likely differences in informational states, goal-directness, workload and opportunity cost implications of sender and receiver roles, we can expect the mechanisms (cognitive and otherwise) which instantiate them to differ as well, quantitatively and qualitatively. This implies that we should not expect their update-responsiveness in any given game to be equal either. Yet the working evolutionary assumption is that senders and receivers update their strategies in an identical manner, modelled using either learning dynamics or replicator dynamics. Call this the evolutionary responsiveness concern.

3 Hedgehog strategies and update asymmetry

The first of these concerns might sound like an argument for abandoning coordination games and moving toward 'conflict of interest' or 'partial conflict of interest' models. However the issue is more specific than this.

The structural responsiveness concern provides parallel motivation to one of Kim Sterelny's (Sterelny, 2012) concerns about Skyrms (2010) use of the Lewis model. Sterelny asks whether the availability of 'third options' on the part of the receiver might undermine the evolution of signalling even when these third options are less valuable than the payoff for successful coordination. As part of a discussion of animal threat responses, he labels this a 'hedgehog' strategy – taking an action which pays off modestly, regardless of the state of the world. To make this concrete, hedgehogs often roll into a ball in response to predators. This is a stark contrast to the more sophisticated behaviour of vervets, who have specific responses to specific threats. Yet the optimal response a vervet takes to one threat – climb a tree when confronted by a leopard – may lead to total disaster when used in response to another threat, such as an eagle. Hedgehogs avoid such outcomes by 'hedging' unconditionally so as to secure a modest payoff. Translated to signalling games, such a gambit may, in many cases, be more attractive than attempting to respond optimally to a signal[2].

[2]It is worth noting here that the 'hedgehog' strategy in this Lewis signalling game is in many ways analogous to the risk dominant 'hare' response in stag hunt games. Playing hare instead of stag allows the agent to avoid disaster, but only guarantees the individual a

This compliments the structural responsiveness concern: receivers (especially) might have other options of value which will stand in competition to those assumed in the standard signalling game. Something like these hedgehog strategies are plausible departures from the idealisation and should be expected on the part of the receiver given a realistic demandingness of the role. The question is whether (as Sterelny suspects) including hedgehog strategies might undermine the robustness of evolution toward signalling systems.

Our second concern pertaining to evolutionary responsiveness parallels a well-known evolutionary hypothesis: the so-called Red Queen effect. In competitive relationships such as predator-prey or parasite-host, the Red Queen hypothesis states that species will be constantly adapting and evolving in response to one another just to "stay in the same place" (Van Valen, 1973). This should also be the case in competitive signalling situations – such as predator-prey signalling systems or courtship displays among conspecifics. Signallers and receivers come to not just update their strategies, but to do so at faster or slower rates depending on the nature of the strategic encounter they are entwined in[3].

It might seem that in David Lewis signalling games (as with games of common interest in general) the Red Queen effect should have no role to play. However any realistic interpretation of the Lewis signalling game makes it plausible to consider asymmetry in evolutionary responsiveness as likely, if not the norm. First, as argued, the precise cognitive mechanisms and procedures employed by senders and receivers are likely to be different. Different systems will admit to different degrees of plasticity and evolvability – and will have a different set of cross-cutting tasks and utilities that will place their own demands upon them. Quick and easy signalling responses will have different pathways of update and adaptation than the (typically) more complex set of systems which appropriate receiver responses require.

The consideration of multiple use or adaptive reuse also makes the Red Queen hypothesis salient: it is wildly implausible that entirely separate cognitive systems would evolve to deal with competitive signalling situations and coordination-style situations. Cognitive structures which underpin sender or receiver behaviour will likely be subject to evolutionary pressures from competitive as well as cooperative situations, and the responsive nimbleness of sender and receiver strategies is therefore not guaranteed to be the same. We should not assume that the evolution of sender and receiver strategies always proceeds at the same pace.

Finally, there is at least some evidence of a basic asymmetry between sender and receiver roles in the literature on great ape communication. For example, Hobaiter and Byrne (2014) stress the great sophistication and flexibility on the receiver side of Chimpanzee gestural communication, while Seyfarth and Cheney (2003) discuss about how greater inferential sophistication on the receiver side is a feature of many primate communication systems. While these findings do

mediocre payoff. Thus the issues and trade-offs associated with the hedgehog strategy are general concerns not confined to just the Lewis signalling games. Thanks to [name redacted for review] for helping us better see this connection.

[3]An example of two groups adapting and evolving at different rates can be found in Richard Dawkin's discussion of his famous Life-Dinner principle (Dawkins and Krebs, 1979). While we expect both predator and prey to adapt to each other, Dawkins claims the prey species will come to evolve at a faster rate than the predator species due to the different selection pressures exerted on both species. Failing to adapt quickly enough for the predator means going hungry for an extra day, while failing to adapt for the prey means death.

not directly support the structural and evolutionary responsiveness concerns, they show that real-life sender and receiver strategies (in our near biological cousins at least) exhibit important differences, suggesting cognitive asymmetries compatible with those concerns.

In summary then, there is reason to consider two structural modifications to the Lewis signalling game as especially salient to the issue of responsiveness: the addition of 'hedgehog' strategies for receivers, and differing rates of change in sender and receiver strategies.

4 The model

The evolutionary model we use as a basis for our analysis is the pure-strategy 2x2x2 David Lewis signalling game, with the two-population discrete-time replicator dynamics.

Exact components of the model include two states of the world (L and R), a world-observing signaller with two possible signals (V1 and V2), and a signal-observing receiver with two possible actions (AL and AR). If the receiver's action matches the state of the world, then both signaller and receiver get a fixed positive success payoff, otherwise their payoff is zero. Signallers and receivers both have four pure strategies available to them (see table 1).

$S1$	Signal V_1 if L and signal V_2 if R
$S2$	Signal V_2 if L and signal V_1 if R
$S3$	Signal V_1 always
$S4$	Signal V_2 always
$S5$	Act A_L if V_1 and act A_R if V_2
$S6$	Act A_R if V_1 and act A_L if V_2
$S7$	Act A_L always
$S8$	Act A_R always

Table 1: Signaller and receiver strategies in the standard 2x2x2 common interest signalling game.

For the evolutionary model, the proportions of the different strategies within sender and receiver populations are initially randomly generated. The fitness of each strategy at a time period t is determined by the composition of the opposing population and the payoff associated with each strategy pairing. The proportion of each strategy at play in the next time period $t+1$ is determined by the standard discrete-time replicator dynamics. For the sender population this is:

$$X_i(t+1) = X_i(t)\frac{F_i}{F_S}$$

where X_i is the ith sender strategy, F_i is the fitness of that strategy and F_S is the average sender strategy fitness. Likewise, for receivers:

$$Y_j(t+1) = Y_j(t)\frac{F_j}{F_R}$$

where Y_j is the jth sender strategy, F_j is the fitness of that strategy and F_R is the average receiver strategy fitness. This is repeated until the populations settle

into an evolutionarily stable arrangement. The update process is deterministic and no randomising or mutations are allowed.

5 Modifications and results

We introduce two novel modifications to this model. First, we add a 'hedgehog' action A_H for the receiver. Second, we allow the rate of generational change of senders and receivers to vary relative to one other. In addition, the bias of nature is also varied, and we investigate the effects these three departures from the Skyrms/Lewis idealisation have on the evolutionary stability of signalling equilibria.

Turning to our first modification, the receiver now has three possible actions upon observing the signal: A_L, A_R, and A_H. As before a success payoff of 1 is received by both players in the case that the receiver plays A_L while the world is in state L, or the receiver plays A_R while the world is in state R. A payoff of zero is received if A_L or A_R is played otherwise. A payoff of H is received unconditionally if the receiver plays A_H, where the value of H is between 0 and 1. The sender has four familiar pure strategies, whereas the receiver now has five (for simplicity we omit conditional strategies involving A_H).

To adapt the earlier forager story, we can imagine the sender and receiver as an egalitarian hunting party, and the game as a situation where the sender remotely observes the location of a valuable prey animal (left or right) and calls out to the receiver. The receiver is initially unable to observe the prey but can choose to go left or go right (catching the prey if they go in the matching direction), or alternatively to abandon the hunt in order to obtain a less valuable resource they do not need help from the sender to acquire (the hedgehog strategy). Varying the prior probability of the world is equivalent to it being in a situation where it is systematically more likely that the prey is to the left or the right.

In the simple unbiased 2x2x2 signalling game, one of the two signalling equilibria is guaranteed to be reached under the replicator dynamics. In our notation, these equilibria are S1-R1 and S2-R2. Increasing the bias of the world (i.e. making L more probable than R or vice versa) will undermine this, with an increasing proportion of populations instead collapsing to pooling equilibria. This will occur when there are initially few conditional signalling strategies in the sender population. In such situations, receivers do best to simply perform the act that is most appropriate for the more likely state of the world. The incentive for senders to adopt a signalling system then disappears and the community is locked into a pooling equilibrium.

Not surprisingly, we found a similar effect with the hedgehog strategy as values of H, the payoff for A_H, becomes significant. The hedgehog strategy R5 is an additional unilateral response, and is able to draw some initial populations away from the signalling equilibria when H is in excess of 0.5 (i.e., the average payoff for 'guessing'). This result, for an unbiased world, is illustrated in Figure 1[4].

[4]Note that the exact range of this effect, including the point at which the effect becomes significant and the y-intercept, are artefacts of the number of world-states and strategies in the model and therefore not general.

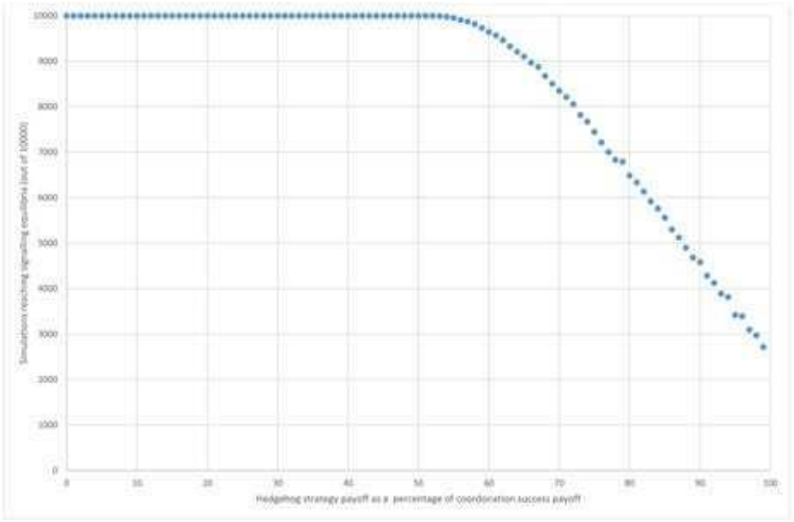


Figure 1: Effect of hedgehog payoff on proportion of signalling equilibria.

We observe a more surprising result when the bias and H are varied in combination. Figure 2 shows the results of varying bias for different values of H. The H = 0 curve has the expected n-shape, with perfect signalling being degraded as world-bias increases away from the mid-point of even bias between L and R. The inclusion of significant (i.e. H ¿ 0.5) hedgehog payoffs decreases signalling at even bias. As nature becomes increasingly biased, however, the proportion of simulations that head to a signalling system does not go down. In fact we observe a 'plateau' followed by a gradual *increase* in the proportion signalling as nature becomes increasingly biased. However, once the bias becomes too extreme, the traditional pooling equilibrium becomes increasingly likely as the payoff associated with simply performing the appropriate act for the more likely state of the world approaches 1. This results in a steep decline in the proportion of simulations that result in signalling systems.

6 Generational asymmetry

We now turn to our second modification of the David Lewis signalling framework in which we introduce a generational asymmetry. We introduced a 'slow-down factor' Z to the replicator dynamics in order control the rate at which sender and receiver populations change over time. Composition of the sender and receiver populations are now governed by the following equations:

$$X_i(t+1) = (1 - Z_S)X_i(t)\frac{F_i}{F_S} + X_i(t)Z_S$$

$$Y_j(t+1) = (1 - Z_R)Y_j(t)\frac{F_j}{F_R} + Y_j(t)Z_R$$

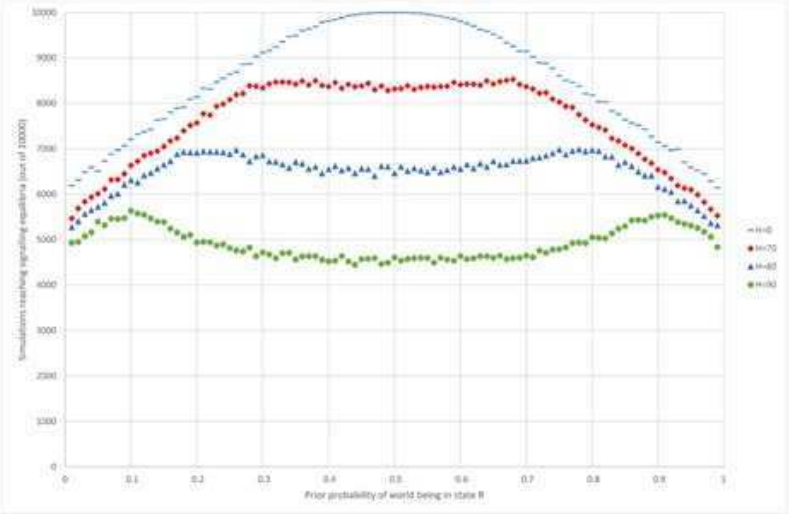


Figure 2: Effect of hedgehog strategy and bias of nature on proportion of signalling equilibria.

Note that when both Z_R and Z_S are zero there is no deviation from the standard replicator dynamics. Rates of changes are slowed as their values increase; for example setting $Z_S = .5$ halves the rate of change for sender strategies. Z_R (alone) being set to 1 means taht the composition of the receiver population would not change over time, and only the sender population would evolve.

The result of introducing this generational asymmetry between senders and receivers is that signalling is more likely when sender strategies evolve faster than receiver strategies. This is illustrated in figure 3, where senders (Z_S) and receivers (Z_R) are slowed down to half and one-tenth speeds (with the other population unaltered) as the bias of nature is varied.

Slowing the evolution of the sender population leads to more pooling because, as before, receivers facing a sender population whose conditional signalling is low will begin to gravitate to the act that matches the more likely state of the world (and the threshold for 'low' is higher at higher bias). This evolutionary trajectory only reverses if conditional signalling increases rapidly enough to tip the fitness balance toward its matching conditional response, before that response is overpowered. Thus signalling becomes quite a remote possibility when bias is high and senders are slow, occurring in less than 10% of simulations for some parameter values. Slowing the evolutionary responsiveness of the receiver population evolves has the opposite effect – as senders will have time to adopt the best separating strategy given the mix of receiver strategies, and the receiver population slowly adjusts and a robust signalling system establishes. By a similar logic, it is easy to see that a quickly evolving sender population also mitigates against the effect of hedgehog strategies.

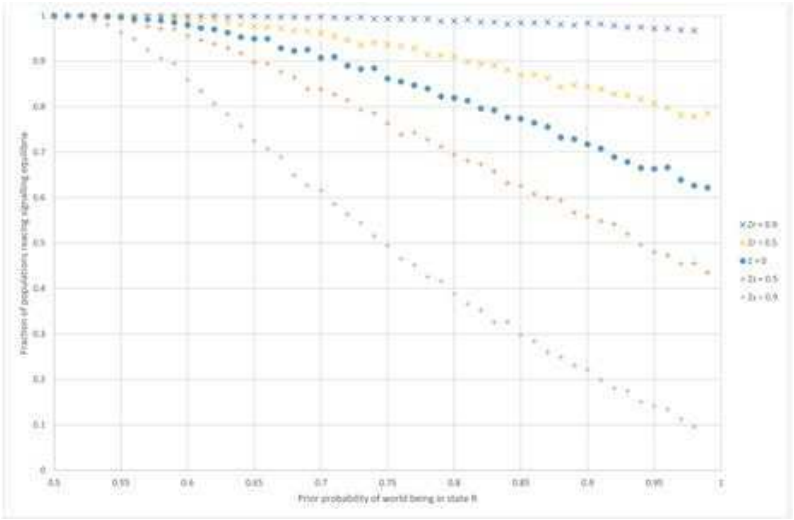


Figure 3: Effect of generational asymmetry and bias of nature on proportion of signalling equilibria.

7 Discussion

We have explored a few well-motivated departures from the highly idealized and simple Lewis signalling game typically considered in the literature. As shown in section 4, breaking the symmetry between senders and receivers often significantly reduces the likelihood that a separating equilibrium emerges. For one, providing receivers with a safe third option which allows them to secure a decent payoff regardless of the state of the world significantly reduces the size of the basin of attraction of the separating equilibrium. Likewise, separating is a remote possibility when receivers outpace senders in the race to adapt.

However the interaction between hedgehog payoffs and bias shows that signalling-undermining effects are not strictly additive. Likewise, the situation is much less bleak when senders evolve at a faster pace than receivers. Interestingly, many scholars in the animal communications literature have noted a similar response asymmetry between sender and receiver in conflict of interest and partial conflict of interest signalling games. For instance, Owren, Rendall, and Ryan (2010) note that senders can easily adapt their signalling behaviour while receivers for the most part have responses to the stimuli produced by senders that are more difficult to change. Thus some have taken to think of signalling as primarily involving the manipulation of receivers by senders.

But this leaves us with an evolutionary puzzle. If there is a conflict of interest between sender and receiver, then what prevents receivers from increasing the speed at which they adapt to the behaviour of the senders? In other words, what explains the absence of an evolutionary arms race between sender and receiver? These are the exact circumstances we would expect the red queen hypothesis to apply. We believe the results of this paper may form the basis of

a novel explanation for this puzzling phenomena. When the interests of sender and receiver are perfectly aligned it is actually in the interest of both parties for the sender population to 'take the lead' and evolve at the faster rate, as doing so ensures the community is more likely to hit upon a mutually beneficial signalling system. When the interests of sender and receiver significantly diverge, however, we would expect this not to be the case since both parties now have reason to adapt at a faster pace than the other.

Yet individuals who routinely interact rarely find themselves playing either common interest or conflict of interest signalling games exclusively. As is well known by any parent, not all signalling interactions between relatives are free of conflict. Likewise, agents whose interests are typically thought to be partially opposed, such as two potential mates, may frequently engage in common interest signalling games in contexts unrelated to mating. The point is that a variety of strategic scenarios can hold between sender and receiver, and there is no principled reason to think all interactions will involve perfect alignment or sizable conflict. If so, then a proportion of signalling interactions between sender and receiver may involve no conflict, a partial conflict, or a full conflict of interest. When the proportion of no or low conflict signalling games is significant, the generational asymmetry result from the previous section may hold to some degree. Both sender and receiver will then profit from the sender population evolving at a faster rate than the receiver population, and receivers do best to limit how responsive they are to senders so as to ensure the emergence of informative signalling systems when their interests do overlap. Thus, while it may appear puzzling as to why a receiver is not more responsive when her interests diverge from that of the sender, this confusion might be resolved when the interaction is put into context.

The robustness analysis considered in this paper has in some sense shown how fragile the evolution of signalling can be. Slightly altering the framework in a sensible fashion leads to significantly different results. While many variants of the baseline Lewis signalling game have been explored by philosophers in recent years, more work is required in order to better assess the prospect of signalling in realistic environments.

8 Acknoweldgements

We thank Kim Sterelny, Ron Planer and the audiences at the Sydney-ANU Philosophy of Biology Workshop and the 2016 Meeting of the Philosophy of Science Association.

9 Bibliography

Bruner, Justin, Cailin O'Connor, Hannah Rubin, and Simon M. Huttegger. 2014. "David Lewis in the Lab: Experimental Results on the Emergence of Meaning." Synthese, September, 1–19. doi:10.1007/s11229-014-0535-x.

Dawkins, R., and J. R. Krebs. 1979. "Arms Races between and within Species." Proceedings of the Royal Society of London B: Biological Sciences 205 (1161): 489–511. doi:10.1098/rspb.1979.0081.

Godfrey-Smith, Peter. 1996. Complexity and the Function of Mind in Nature. Cambridge; New York: Cambridge University Press.

Godfrey-Smith, Peter, and Manolo Martínez. 2013. "Communication and Common Interest." PLoS Comput Biol 9 (11): e1003282. doi:10.1371/journal.pcbi.1003282.

Hobaiter, Catherine, and Richard W. Byrne. 2014. "The Meanings of Chimpanzee Gestures." Current Biology 24 (14): 1596–1600. doi:10.1016/j.cub.2014.05.066.

Huttegger, Simon M. 2007. "Evolution and the Explanation of Meaning*." Philosophy of Science 74 (1): 1–27. Lewis, David K. 1969. Convention: A Philosophical Study,. Cambridge, MA: Harvard University Press.

Martınez, Manolo, and Peter Godfrey-Smith. 2015. "Common Interest and Signaling Games: A Dynamic Analysis." http://petergodfreysmith.com/wp-content/uploads/2013/06/Martinez-GS-paper2-Dynamic-Preprint.pdf.

Owren, Michael J., Drew Rendall, and Michael J. Ryan. 2010. "Redefining Animal Signaling: Influence versus Information in Communication." Biology and Philosophy 25 (5): 755–80. doi:10.1007/s10539-010-9224-4.

Pawlowitsch, Christina. 2008. "Why Evolution Does Not Always Lead to an Optimal Signaling System." Games and Economic Behavior 63 (1): 203–26. doi:10.1016/j.geb.2007.08.009.

Seyfarth, Robert M., and Dorothy L. Cheney. 2003. "Signalers and Receivers in Animal Communication." Annual Review of Psychology 54 (1): 145–73. doi:10.1146/annurev.psych.54.101601.145121.

Skyrms, Brian. 1996. Evolution of the Social Contract. Cambridge University Press. ———. 2010. Signals: Evolution, Learning, and Information. Oxford; New York: Oxford University Press.

Sterelny, Kim. 2012. "A Glass Half-Full: Brian Skyrms's Signals." Economics and Philosophy 28 (01): 73–86. doi:10.1017/S0266267112000120.

Van Valen, Leigh. 1973. "A New Evolutionary Law." Evolutionary Theory 1 (1-30). http://tmtfree.hd.free.fr/albums/files/TMTisFree/Documents/Biology/A

Experimental Individuation and Retail Arguments

Ruey-Lin Chen
Department of Philosophy, National Chung Cheng University, Taiwan

Jonathon Hricko
Education Center for Humanities and Social Sciences, National Yang-Ming University, Taiwan

Abstract: Magnus and Callender (2004) argue that we ought to focus on retail arguments, which are arguments regarding the existence of particular kinds of theoretical entities, as opposed to theoretical entities in general. However, *scientists* are the ones who put forward retail arguments, and it's unclear how *philosophers* can engage with such arguments. We argue that philosophers can engage with retail arguments by providing criteria that they must satisfy in order to demonstrate the existence of theoretical entities. We put forward experimental individuation as such a criterion—when scientists experimentally individuate an entity, a realist conclusion about that entity is warranted.

Word Count: 4983

1. Introduction

Magnus and Callender argue that we ought to abandon "wholesale arguments," which are "arguments about all or most of the entities posited in our best scientific theories" (2004, 321). Instead, we ought to embrace "retail arguments," which are "arguments about specific kinds of things such as neutrinos, for instance" (2004, 321). This shift in focus rules out standard scientific realism as well as various antirealist positions, and in Section 2, we'll argue that Magnus and Callender's position is preferable to these other positions.

However, we recognize that philosophers who choose to abandon wholesale arguments in favor of retail arguments face a potential problem. Dicken (2013) has argued that such philosophers will merely end up repeating the retail arguments that scientists offer. In that case, the turn to retail arguments may entail that no distinctively philosophical work remains to be done. In Section 3, we'll argue that this is not the case. Not all retail arguments successfully demonstrate the existence of theoretical entities, and it can take some philosophical work to distinguish the ones that do from the ones that don't.

In Section 4, we'll put forward a criterion for doing so, which we take from Chen's (2016) work on experimental individuation. Chen suggests that "[i]f a scientist can realize the individuality of an object in a particular experiment, then she has provided the strongest evidence ... to warrant the reality of the object" (2016, 365). We'll argue that retail arguments that demonstrate the experimental individuation of a theoretical entity succeed in showing that realism about that entity is warranted.

We'll draw on three examples throughout the paper: Lavoisier's oxygen theory of acidity, J. J. Thomson's work on cathode rays, and Davy's discovery of potassium. We'll conclude, in Section 5, by applying our criterion to these three cases, with the result that the upshot of a retail argument can be either realism, antirealism, or skepticism regarding the existence of a particular kind of theoretical entity.

2. The Turn to Retail Arguments

We'll now introduce Magnus and Callender's position in a bit more detail, and indicate why we take it to be preferable to standard scientific realism (SSR) and antirealism. SSR is a position regarding theories in general—the success of our best theories warrants the claim that they are at least approximately true, as well as the claim that the theoretical entities that they posit exist. Antirealist positions come in a number of different forms, but they all typically endorse claims about theories in general, and deny that success warrants the two claims endorsed by proponents of SSR.

According to Magnus and Callender, there is something that all of these positions have in common, namely, their proponents attempt to support these positions by engaging in wholesale arguments. They focus on two examples of such arguments. First of all, there is the no-miracles argument, according to which the success of our best theories would be a miracle if those theories weren't at least approximately true. Secondly, there is the pessimistic meta-induction, which uses past successful-but-false theories as an inductive basis for concluding that our current successful theories are false as well. The no-miracles argument is taken to support "[w]holesale realism," which "seeks to explain

the success of science in general"; and the pessimistic meta-induction is taken to support "wholesale anti-realism," which "seeks to explain the history of science in general" (2004, 321). However, Magnus and Callender argue that these arguments, and wholesale arguments in general, ought to be abandoned. This is because they embody the base rate fallacy, since they don't take into account the base rate probability of any successful theory being true or false. For this reason, they maintain that wholesale realism and wholesale antirealism ought to be abandoned as well.

Magnus and Callender propose that we ought to replace wholesale arguments with retail arguments. Unlike wholesale arguments, the scope of a retail argument is restricted to a particular theory and/or a particular kind of theoretical entity. By shifting the focus from theories in general to theories in particular, philosophers can *dissolve* the traditional realism debate, with the result that "realism and anti-realism are options to be exercised sometimes here and sometimes there" (2004, 337). This, in turn, opens up the possibility that "[t]here may be good reasons to be a realist about neutrinos, an anti-realist about top quarks, and so on" (2004, 333).

In order to show why this possibility represents an improvement over SSR and antirealism, we'll now consider a case from the history of chemistry. This case concerns the composition of hydrochloric acid. Scheele was the first to decompose this acid, which he called "acid of salt," and he identified its constituent substances as phlogiston and "dephlogisticated acid of salt" (1774/1931). However, it was a matter of some controversy whether he had succeeded in decomposing hydrochloric acid. According to Lavoisier's oxygen theory of acidity, all acids are composed of oxygen (the principle of acidity) and a radical, which can be either a simple substance or a compound (1789/1965,

65, 115). Neither Scheele nor any other chemist had been able to extract the oxygen from hydrochloric acid, which Lavoisier called "muriatic acid." And so Lavoisier held that it remained undecomposed, and, in accordance with his theory, he hypothesized that it must contain oxygen combined with what he called "the muriatic radical" (1789/1965, 71-72). As for Scheele's dephlogisticated acid of salt, Lavoisier held that it is a compound of muriatic acid and oxygen, which he called "oxygenated muriatic acid" (1789/1965, 73). Some years later, Davy argued that Scheele was correct, while Lavoisier was in error (1810, 236-37). On Davy's view, muriatic acid is composed of hydrogen and what he calls "oxymuriatic acid," which is what Lavoisier called "oxygenated muriatic acid," and what Scheele called "dephlogisticated acid of salt." Davy later went on to argue for the elementary nature of this latter substance, and proposed a new name for it: "Chlorine" (1811, 32). His approval of Scheele stems from the fact that Davy, like a number of latter-day phlogiston theorists, identified hydrogen with phlogiston.[1] And the claim that hydrochloric acid is made up of hydrogen and dephlogisticated acid of salt, even if terminologically problematic, is essentially correct. Lavoisier, however, was in error since this acid contains no oxygen, thus falsifying his oxygen theory of acidity.

Proponents of SSR, impressed by narratives of the Chemical Revolution according to which Lavoisier's oxygen theory defeated the phlogiston theory, are often explicit that their realism applies to the oxygen theory but not to the phlogiston theory.[2] But in that case, SSR entails the implausible conclusion that Lavoisier's muriatic radical exists, while Scheele's dephlogisticated acid of salt does not. It seems much better to

[1] See, e.g., Kirwan (1789, 4-5).

[2] See, e.g., Hardin and Rosenberg (1982, 610) and Psillos (1999, 291).

conclude that Lavoisier's muriatic radical doesn't exist, while Scheele's dephlogisticated acid of salt does.

Antirealism, at least of the Kuhnian variety, fares no better. Those influenced by Kuhn's (1962/1996) views regarding incommensurability would claim that theoretical entities conceptualized by rival theories should be treated as different entities. However, chemists working in the late eighteenth and early nineteenth centuries shared a set of operations for producing the substance that was variously known as dephlogisticated acid of salt, oxymuriatic acid, and chlorine. It's therefore implausible to maintain that, in light of the fact that these chemists held different theories, they were working with distinct theoretical entities. A trans-theoretical view of the substance that came to be known as chlorine is therefore preferable.

By abandoning wholesale arguments in favor of retail arguments, we can sidestep these difficulties, and simply adopt realism about chlorine (whatever it was called and however it was conceptualized) and antirealism about Lavoisier's muriatic radical. That said, by trading wholesale arguments for retail arguments, we face another difficulty, to which we'll now turn.

3. Can Philosophers Engage with Retail Arguments?

Dicken (2013) has objected that those who abandon wholesale arguments in favor of retail arguments face a serious difficulty. In short, once one does so, it's not clear that any "distinctively philosophical" issues remain to be addressed (2013, 564). Scientists are generally the ones who put forward retail arguments. And if the turn to retail arguments

amounts to merely repeating arguments scientists have offered first, then perhaps nothing distinctively philosophical remains to be done. Our goal in the remainder of the paper is to provide a way of engaging with retail arguments that is distinctively philosophical, and to thereby answer Dicken's objection.

We'll start by considering how scientists demonstrate the existence of theoretical entities, and so we'll now introduce another case from the history of science. This case concerns Thomson's work on cathode rays and his determination of the mass-to-charge ratio (m/e) of the electron. According to the official website of the Nobel Prize, it was because of this work that Thomson "received the Nobel Prize in 1906 for the discovery of the electron, the first elementary particle."[3] Thomson (1897, 1906/1967) hypothesized that cathode rays are currents of "carriers of negative electricity" or "corpuscles"—what we now know as electrons.[4] His hypothesis was not only about the nature of cathode rays, but also about the interaction among cathode rays and other theoretical entities such as electrostatic fields and electrons. In order to determine the mass-to-charge ratio, he measured the deflection of cathode rays passing through an electrostatic field, the strength of the electrostatic field, and other related magnitudes. He interpreted the value that he obtained for m/e in light of his hypothesis, and his experimental results confirmed that hypothesis.

[3] Retrieved January 27, 2016 from http://www.nobelprize.org/educational/physics/vacuum/experiment-1.html. See also Harré (2002) and Whittaker (1989).

[4] For the identification of Thomson's carriers with electrons, see the reprint of Thomson (1897) in Magie (1969), in which Magie makes the identification.

However, one might ask how it's possible to infer from Thomson's experimental confirmation of his hypothesis to the claim that he had thereby demonstrated the existence of the electron. Philosophers can engage with such a question. And regardless of the answers they provide, they must at least defend those answers by invoking some kind of criterion for concluding that the evidence that scientists have offered does or does not constitute a demonstration of the existence of a given entity. To take one example of such a criterion, Hacking (1983, 23) suggests manipulation: "if you can spray them then they are real." While Thomson manipulated cathode rays, he did not manipulate electrons, and so, according to Hacking's criterion, Thomson did not offer evidence strong enough to demonstrate the existence of electrons.

The important point, for our purposes, is that providing a criterion for granting the reality of a theoretical entity, and determining whether the evidence that scientists have offered satisfies that criterion, constitutes a way for philosophers to engage with retail arguments. Scientists may be the ones who initially put forward retail arguments. But it is a distinctively philosophical task to determine a criterion that can distinguish those retail arguments that demonstrate the existence of a theoretical entity from those that do not. We thus have a way of answering Dicken's objection, provided that, by invoking such a criterion, we are not thereby turning back to wholesale arguments. In the next section, we'll introduce our criterion and argue that applying it does not amount to a wholesale argument.

4. Ontological Commitment and Experimental Individuation

Our proposed criterion for granting the reality of theoretical entities is experimental individuation. A retail argument that demonstrates the experimental individuation of an entity is a good argument for realism about that entity.

Individuation and ontological commitment are connected. When scientists are ontologically committed to the theoretical entities that they posit, this commitment involves not just a belief that the entities exist, but also a responsibility to demonstrate their existence. Demonstrating the existence of a posited entity requires scientists to find an individual instance or sample of that entity, and if a scientist posits a theoretical entity without individuating it, then her ontological commitment is empty.

How do scientists individuate theoretical entities? Answering this question requires us to distinguish *theoretical individuation* from *experimental individuation*. Scientists theoretically individuate an entity if, in the course of theorizing, they describe a set of properties and behaviors of a posited entity by which they can identify it and distinguish it from other entities. However, these descriptions by which scientists theoretically individuate entities require evidence. Scientists can offer evidence for the existence of a theoretical entity if they produce an instance or sample of such an entity by performing an experiment. In doing so, they individuate an entity experimentally.[5]

The relationship between theoretical individuation and experimental individuation is much the same as the relationship between theory and experiment more generally.

[5] Scientists may also individuate an entity *observationally*, by observing an instance or sample of such an entity. Since observation is itself a complex issue, and since participants in the realism debate rarely question the existence of entities that scientists have observed, we will not discuss observational individuation here.

Various worries about the theory-ladenness of experimentation are relevant here. If a theoretical hypothesis yields a prediction regarding some experimental result, the result may be interpreted in light of the hypothesis. Moreover, since a theoretical hypothesis may involve two or more theoretical entities and their interactions, it can be difficult to show that an experiment produces an instance or sample of the target entity, i.e., that it experimentally individuates that entity. And it can be difficult to judge whether an experiment produces a real individual, as opposed to a mere phenomenon that results from experimental apparatuses and their interactions with experimented objects. For these reasons, a criterion of experimental individuation that is sufficiently independent of theoretical interpretation is needed.

Is there such a criterion for experimental individuation? One candidate is Hacking's manipulation criterion, which we mentioned in Section 3. However, since experimenters can manipulate not just real individuals, but also mere phenomena, manipulation cannot singly serve as the criterion of experimental individuation. Chen (2016) takes Hacking's criterion of manipulation, along with two other criteria, namely, separation and maintenance of structural unity, as jointly constituting a necessary and sufficient condition for the experimental individuation of a theoretical entity. In short, experiments that produce individuals are experiments that separate individuals from their surrounding environment, manipulate them, and maintain their structural unity throughout the process. Importantly, Chen's further conditions ensure that the manipulated object is a real individual as opposed to a mere phenomenon. We take Chen's criteria to offer a satisfactory account of experimental individuation. In Section 5, we'll illustrate his criteria in terms of three retail arguments from the history of science,

and thereby provide some support for our claim that his criteria are satisfactory.

For now, we wish to emphasize two points. First of all, experimental individuation is our proposed criterion for determining whether a retail argument successfully demonstrates the existence of some theoretical entity—it succeeds if it demonstrates the experimental individuation of that entity. Secondly, Chen's three criteria provide an adequate account of what experimental individuation requires.

Before moving on, we'll discuss two potential problems with this proposal. First of all, some theoretical entities, like the chemical substances named by mass terms like 'water,' 'phlogiston,' and 'oxygen,' are paradigm cases of non-individuals. It's therefore not immediately obvious how we can appeal to the notion of experimental individuation when it comes to such entities. We propose to do so by considering the experimental individuation of *samples* of such substances, as we'll illustrate in Section 5.1, in terms of Davy's discovery of potassium. Since samples count as individuals, our criterion is applicable to cases involving non-individuals like chemical substances.

Secondly, there's the issue as to whether the application of our criterion amounts to a kind of wholesale argument. Whether a given retail argument demonstrates the experimental individuation of some theoretical entity is a local matter, grounded in the details of that argument. In contrast, wholesale arguments are not grounded in such local matters. Instead, they rely on claims regarding populations of theories in general, and it is for this reason that they embody the base rate fallacy. We've consciously avoided reasoning that may lead to the base rate fallacy. For example, we haven't argued that the success of our best theories would be a miracle unless the entities they posit can be experimentally individuated. For these reasons, the application of our criterion to retail

arguments does not amount to a kind of wholesale argument. And in that case, we've provided a way of answering Dicken's objection, since our criterion provides a way for philosophers to engage with retail arguments.

5. Application of the Criterion to Three Retail Arguments

Our goal at this point is to show how one can use the criterion we've proposed in order to engage with retail arguments regarding the existence of particular kinds of theoretical entities. We'll discuss three cases: Davy's potassium, Lavoisier's muriatic radical, and Thomson's electron.

5.1 A Realist Conclusion Regarding Davy's Potassium

To begin with, we'll argue that Davy demonstrates the experimental individuation of potassium, and thereby provides us with a successful retail argument for realism about that substance.

Davy first isolated potassium by decomposing potash, which he did by means of electrolysis (1808, 4-5). He was the first to decompose potash, though for some time, chemists suspected it to be a compound.[6] Davy acted on a small piece of moistened potash with a Voltaic battery. As a result, at the negative surface of the battery Davy observed the appearance of "small globules having a high metallic lustre, and being precisely similar in visible characters to quicksilver" (1808, 5). In the lecture in which he

[6] See, e.g., Lavoisier (1965/1789, 156).

reports these results, Davy goes on to write: "These globules, numerous experiments soon shewed to be the substance I was in search of, and a peculiar inflammable principle the basis of potash" (1808, 5). And later in the lecture, he proposes the name "Potasium [sic]" for the basis of potash (1808, 32).

While this experiment, on its own, does not demonstrate the experimental individuation of a sample of potassium, subsequent experiments that Davy conducted do, and he shows that potassium satisfies all three of Chen's criteria. First of all, there is Chen's separation condition: scientists must separate the entities that they produce "from their environments" (2016, 348), and "from the experimental instruments that may have helped produce [them]" (2016, 365). In order to determine whether his results depended on the platinum instruments that he used, Davy performed a number of experiments using a variety of other materials, including copper, silver, and gold (1808, 5). And in order to determine whether his results depended on the fact that he conducted his experiments in the open atmosphere, he performed similar experiments in a vacuum (1808, 5). In all of these cases, he obtained the same results. These experiments collectively show that Davy had separated potassium from its surrounding environment (including the atmosphere and the other components of potash), and from the instruments that he used, thereby satisfying Chen's separation condition.

Secondly, there is Chen's condition regarding the maintenance of structural unity. Chen understands structural unity as the idea that "the components of an individual are structured into a whole in some specific manner" (2016, 358). Davy encountered a number of difficulties when it came to maintaining the structural unity of the globules of potassium that he had produced because "they acted more or less upon almost every body

to which they were exposed" (1808, 10). One of the first things Davy notes about the globules is that they did not last long—the ones that did not explode immediately after forming soon lost their metallic luster and became "covered by a white film" (1808, 5). Davy identifies this film as pure potash, and explains how it attracts moisture from the atmosphere, converting the globule into a saturated solution of potash (1808, 7). Eventually, Davy discovered one substance on which potassium did not have much of an effect, namely, recently distilled naphtha (1808, 10). He used that fluid to preserve globules of potassium, and he was able to examine the properties of potassium in the atmosphere by covering the globules with a thin film of naphtha. This method allowed Davy to maintain the structural unity of potassium, thus satisfying Chen's condition.

Thirdly, there is Chen's manipulation condition. Chen understands this condition in terms of the "instrumental use" of an object "to investigate other phenomena of nature" (2016, 358). Towards the end of the lecture in which he reports the electrolytic decomposition of potash, Davy conjectures that the globules of potassium he isolated "will undoubtedly prove powerful agents for analysis; and having an affinity for oxygene [sic] stronger than any other known substances, they may possibly supersede the application of electricity to some of the undecompounded bodies" (1808, 44). Making good on this conjecture would amount to showing that chemists can use potassium to decompose previously undecomposed substances, thereby satisfying Chen's manipulation condition. And in the following year, Davy made good on this conjecture by using potassium to extract the oxygen from a previously undecomposed substance, namely, boracic acid, thereby decomposing it (1809, 76-77).

In sum, Davy shows that samples of potassium satisfy all three of Chen's criteria.

And by demonstrating the experimental individuation of these samples, Davy presents us with a successful retail argument for realism about potassium.

5.2 An Antirealist Conclusion Regarding Lavoisier's Muriatic Radical

We'll now argue that Davy shows why the experimental individuation of Lavoisier's muriatic radical is not possible, and thereby provides us with a successful retail argument for antirealism about Lavoisier's radical.

As we discussed in Section 2, Lavoisier hypothesized that hydrochloric acid, which he called muriatic acid, is composed of oxygen and a hypothetical substance that he called the muriatic radical. He thereby theoretically individuated the muriatic radical as that substance which combines with oxygen to form muriatic acid, which, in turn, is converted into oxymuriatic acid (i.e., chlorine) by means of combining with even more oxygen. But as we emphasized in Section 4, theoretical individuation is a mere belief, and beliefs require evidence.

Davy (1810, 235-36) provides a retail argument that demonstrates that the experimental individuation of Lavoisier's radical is not possible. He emphasizes the results of various experiments that he and other chemists performed, which show that oxymuriatic acid combines with hydrogen to form muriatic acid. And he goes on to discuss those experiments that seem to show the decomposition of oxymuriatic acid into oxygen and muriatic acid. Davy observes that in these experiments, water is always present. And he concludes that the oxygen that such experiments produce results from the decomposition of the water, not from the decomposition of oxymuriatic acid, which has

not been demonstrated. If oxymuriatic acid doesn't contain oxygen, and muriatic acid contains oxymuriatic acid and hydrogen, then muriatic acid doesn't contain oxygen either. To adopt Davy's later terminology, the only components of muriatic acid are hydrogen and chlorine. Experimentally individuating the muriatic radical would involve separating it from the oxygen with which it combines to form muriatic acid and oxymuriatic acid. And since Davy showed that this is not possible, he gives us a successful retail argument for antirealism about Lavoisier's radical.

5.3 A Skeptical Conclusion Regarding Thomson's Electron

Finally, we'll argue that Thomson neither demonstrates the experimental individuation of the electron, nor shows that it is impossible. Hence, we have an example of an inconclusive retail argument. The proper response to such an argument is skepticism regarding the entity in question, at least until there is a conclusive retail argument regarding the existence of that entity.

Thomson (1897) designed a new type of cathode ray tube (figure 1) to perform a deflection experiment.

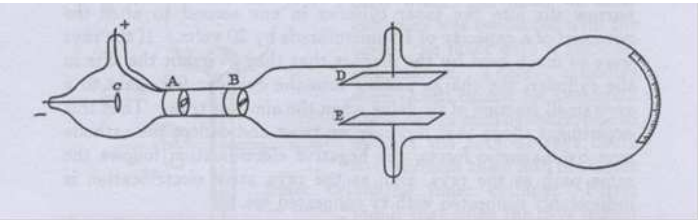


Figure 1. Thomson's cathode ray tube in 1897. Reproduced from Thomson 1969, 586.

This tube contains a cathode *C*, a cylindrical anode *A* with a slit, a cylindrical metal ring *B* with a slit, and a pair of plates *D* and *E* that produce an electrostatic field. A cathode ray is produced when the cathode discharges, and the ray passes through the slits in *A* and *B* before passing through the electrostatic field produced by *D* and *E*. Thomson's goal was to determine whether the ray would be deflected in the field, and to thereby determine the composition of cathode rays. The basic idea was that, if cathode rays were made of ethereal waves, the rays would not be deflected by an electrostatic field; if, however, the rays were made up of negatively electrified bodies, then the rays would be deflected by an electrostatic field.

Thomson's thought was that a cathode would produce both electric currents and cathode rays when discharging, and that, in order to determine the composition of cathode rays, it would be necessary to eliminate the electric currents and experiment with purified cathode rays. Purification is the function of the cylindrical metal ring *B*, which absorbs the electric currents leaked from *A* and thus ensures that the ray passing through *B* is pure. Thomson found that the purified cathode ray was deflected when it passed between the plates *D* and *E*, thus confirming that cathode rays are made up of negatively electrified bodies.

While Thomson satisfies Chen's criteria when it comes to cathode rays, he didn't thereby experimentally individuate the electrons that make them up. Thomson succeeded in *separating* cathode rays from currents; purifying them with the metal ring *B*, and thus *maintaining their structural unity*; and *manipulating* them by deflecting them with an electrostatic field. According to Chen's criteria, one can say that Thomson experimentally individuated cathode rays and demonstrated that they are currents of

negative electricity. But Thomson *presupposed* rather than demonstrated that the currents consist of electrons. He did not demonstrate the existence of electrons, because he did not experimentally individuate them. Hence, the proper response to the retail argument that Thomson gives us is neither realism nor antirealism, but rather skepticism regarding the existence of electrons, at least until there is a conclusive retail argument.

6. Conclusion

Our goal in this paper has been to provide a way for philosophers to engage with retail arguments, and thereby show that, even if we dissolve the traditional realism debate, there is still philosophical work to be done. We've put forward the criterion of experimental individuation in order to determine whether a given retail argument demonstrates the existence of a particular kind of theoretical entity. And we've applied that criterion to three cases, with the result that the upshot of a retail argument can be either realism, antirealism, or skepticism regarding the existence of a particular kind of theoretical entity.

References

Chen, Ruey-Lin (2016). "Experimental Realization of Individuality." In *Individuals Across the Sciences*, ed. Thomas Pradeu and Alexandre Guay, 348-70. New York: Oxford University Press.

Davy, Humphry (1808). "The Bakerian Lecture [for 1807], on Some New Phenomena of Chemical Changes Produced by Electricity, Particularly the Decomposition of the Fixed Alkalies, and the Exhibition of the New Substances Which Constitute Their Bases; And on the General Nature of Alkaline Bodies." *Philosophical Transactions of the Royal Society of London* 98: 1-44.

— (1809). "The Bakerian Lecture [for 1808]: An Account of Some New Analytical Researches on the Nature of Certain Bodies, Particularly the Alkalies, Phosphorus, Sulphur, Carbonaceous Matter, and the Acids Hitherto Undecompounded; With Some General Observations on Chemical Theory." *Philosophical Transactions of the Royal Society of London* 99: 39-104.

— (1810). Researches on the Oxymuriatic Acid, Its Nature and Combinations; And on the Elements of the Muriatic Acid. With Some Experiments on Sulphur and Phosphorus, Made in the Laboratory of the Royal Institution. *Philosophical Transactions of the Royal Society of London* 100: 231-57.

— (1811). The Bakerian Lecture [for 1810]: On Some of the Combinations of Oxymuriatic Gas and Oxygene, and on the Chemical Relations of These Principles, to Inflammable Bodies. *Philosophical Transactions of the Royal Society of London* 101: 1-35.

Dicken, Paul (2013). "Normativity, the Base-Rate Fallacy, and Some Problems for Retail Realism." *Studies In History and Philosophy of Science* 44(4): 563-70.

Hacking, Ian (1983). *Representing and Intervening*. Cambridge: Cambridge University Press.

Hardin, Clyde L. and Alexander Rosenberg (1982). In Defense of Convergent Realism. *Philosophy of Science* 49(4): 604-15.

Harré, Rom (2002). *Great Scientific Experiments: Twenty Experiments that Changed our View of the World*. New York: Dover.

Kirwan, Richard (1789). *An Essay on Phlogiston and the Constitution of Acids*. 2nd ed. London: J. Johnson.

Kuhn, Thomas S. (1962/1996). *The Structure of Scientific Revolutions*. 3rd ed. Chicago: University of Chicago Press.

Lavoisier, Antoine Laurent (1789/1965). *Elements of Chemistry*. New York: Dover.

Magie, William Francis, ed. (1969). *A Source Book in Physics*. Cambridge, Mass.: Harvard University Press.

Magnus, P. D. and Craig Callender (2004). "Realist Ennui and the Base Rate Fallacy." *Philosophy of Science* 71(3): 320-38.

Psillos, Stathis (1999). *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Scheele, Carl Wilhelm (1774/1931). On Manganese or Magnesia; and Its Properties. In *The Collected Papers of Charles Wilhelm Scheele, Translated from the Swedish and German Originals by Leonard Dobbin*, 17-49. London: G. Bell and Sons.

Thomson, Joseph John (1897). "Cathode Rays." *Philosophical Magazine, Fifth Series* 44: 293-316.

— (1906/1967). "Carriers of Negative Electricity. Nobel Lecture, December 11, 1906." In *Nobel Lectures: Physics, 1901-1921*, 145-53. Amsterdam: Elsevier Press.

— (1969). "The Electron." In Magie 1969, 583-97.

Whittaker, Edmund Taylor (1989). *A History of the Theories of Aether and Electricity*. New York: Dover.

Crash Testing an Engineering Framework in Neuroscience: Does the Idea of Robustness Break Down?[1]

ABSTRACT

In this paper I discuss the concept of *robustness* in neuroscience. Various mechanisms for making systems robust have been discussed across biology and neuroscience (e.g. redundancy and fail-safes). Many of these notions originate from engineering. I argue that concepts borrowed from engineering aid neuroscientists in (1) operationalizing robustness; (2) formulating hypotheses about mechanisms for robustness; and (3) quantifying robustness. Furthermore, I argue that the significant disanalogies between brains and engineered artefacts raise important questions about the applicability of the engineering framework. I argue that the use of such concepts should be understood as a kind of simplifying idealization.

"The brain is a physical device that performs specific functions; therefore, its design must obey general principles of engineering."
Sterling and Laughlin (2015:xv)

1. INTRODUCTION

In this paper I discuss a cluster of issues around the understanding of *robustness* in neuroscience. Systems biologist, Hiroaki Kitano defines

[1] M. Chirimuuta. History & Philosophy of Science, University of Pittsburgh. mac289@pitt.edu. Accepted for presentation at the 2016 Philosophy of Science Association meeting and publication of the proceedings in *Philosophy of Science.*

robustness as, "a property that allows a system to maintain its functions against internal and external perturbations" (Kitano 2004, p.826). According to this definition, in order to determine whether or not a system is robust, one must specify its function, and also specify the kinds of perturbation it faces. Empirically determinable questions then follow about how exactly the system achieves its robustness. Various means for making systems robust have been discussed across biology and neuroscience: copy redundancy, fail-safes, degeneracy, modularity, passive reserve, active compensation, plasticity, decoupling, and feedback (see Figure 1). It is obvious, but still worth emphasising, that most of these notions originate from engineering.

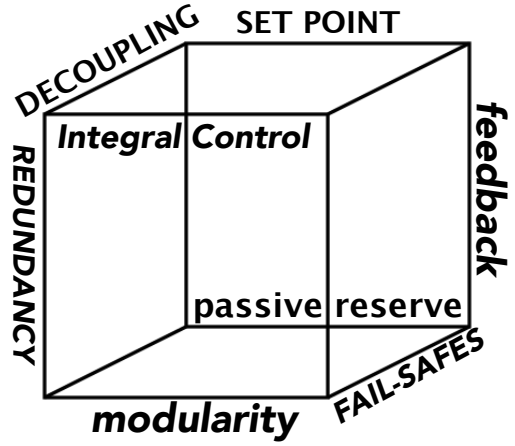


FIGURE 1. The Engineering Framework for Robustness. A set of terms originating from engineering and control theory, which are applied to biological systems to explain how they achieve robust performance.

In Section 2 of this paper I argue that the framework of concepts borrowed from engineering aids neuroscientists in (1) operationalizing robustness by specifying functions of the system and determining possible sources of

perturbation; (2) formulating hypotheses about means for the system to achieve robustness; and (3) showing how robustness may be precisely quantified. This will be shown with examples of neuroscientific research which aims to measure robustness in a retinal circuit (Sterling and Freed 2007), in the motor cortex (Svoboda 2015), and to develop models of homeostatic control (Davis 2006, O'Leary 2014).

In Section 3 I argue that the use of the engineering framework in neuroscience gets stretched, perhaps to breaking point, when applied to systems where (1) there is no principled distinction between processes for robustness and processes which continually maintain the life of the cell; (2) where perturbations are a regular occurrence rather than anomalous events; and (3) where one should not conceive of the system as seeking to maintain a steady state. This point will be illustrated through examination of some recent work from Eve Marder's laboratory, one of the key centres for research on robustness in neuroscience.

I will argue that the limitations of the engineering notions are put into stark relief when one examines neural systems through the lens of the process approach to biology (Dupré 2012). The engineering perspective, to the extent that it treats biological systems as pre-specified objects with fixed functions, misses many of the features that make robust biological systems fascinating and which are highlighted by the process view.

In Section 4 I will consider if it is necessary to re-engineer the concepts of robustness to be more in line with the dynamicism of biological systems; or alternatively, if we should accept the engineering perspective as it is, as one amongst many idealizing and simplifying heuristics for understanding complex systems like the brain.

2. PUTTING THE ENGINEERING FRAMEWORK TO USE

The robustness of the brain is one of its many extraordinary attributes. By this I mean the fact that brains can undergo moderately severe external perturbations while still maintaining approximately normal function. Obviously, robustness has its limits and the brain's characteristic patterns of resilience and fragility are an important target of research (Sporns 2010, chap. 10). In order to investigate robustness it is necessary first to specify what sorts of perturbations the system is robust to, and then to quantify how robust it actually is. Explanations of robustness can be developed by testing hypotheses concerning the exact mechanisms by which robust performance is achieved. The engineering framework can be put to effective use in each of these processes.

For example, Sterling and Freed (2007) pose the question of how robust the retinal circuit is. They define robustness as the factor by which intrinsic capacity exceeds normal demand, which is the engineer's notion of margin of safety (p.563). The idea can be illustrated through their comparison with bridge design. An engineer designing a road bridge will consider both the anticipated normal demand (e.g. commuter traffic) as well as the unusual demands that might occasionally be placed on the bridge (e.g. the passage of a 30 ton military vehicle). The unusual demand can be thought of as a "perturbation" in Kitano's terms. A robust design will ensure that the system does not break when pushed beyond normal conditions. For a bridge this can be achieved with passive reserve (using thicker steel than is needed under normal conditions) and redundancy (including additional beams so that there are back-up structures if any parts are compromised).

Sterling and Freed take the bridge case to be analogous to the retinal circuit. Normal demand, for the retina, is the intensity of illumination that the eye will encounter under naturalistic stimulation conditions. The safety factor is calculated by experimental determination of the maximum illumination level under which neurons in the retina can maintain their ability to signal to downstream neurons. Sterling and Freed (2007, p.570) report that,

> "across successive stages in this neural circuit, safety factors are on the order of 2–10. Thus, they resemble those in other tissues and systems. Their similarity across stages also accords with the principle of symmorphosis—that efficient design matches capacities across stages that are functionally coupled…."

Sterling and Freed's explanation of robustness depends on the notion of passive reserve. For photoreceptor neurons, this is calculated as the number of vesicles of neurotransmitter available in their synapse for continuous signalling at high-rates without restocking of the vesicles (p.565-6). In arriving at their conclusion about retinal safety margins, they argue that there are at least twice as many vesicles as needed under normal stimulation conditions. In this case we have seen that a design approach borrowed from civil engineering plays a clear and striking role in these neuroscientist's definition, operationalization and explanation of robustness in the retina.

Another example comes from Davis's (2006) review of work on homeostatic regulation[2] in the nervous system. As he writes:

[2] Note that Davis makes a conceptual distinction between robust properties and properties under homeostatic control: "In general, robustness describes a system with a reproducible output, whereas homeostasis refers to a system with a constant output" (2006, p.308). I will ignore this difference for the purposes of the paper since homeostatic systems conform to Kitano's general definition robust systems.

> "Homeostatic control systems are best understood in engineering theory, where they are routinely implemented in systems such as aircraft flight control. Recently, biological signaling systems have been analyzed with the tools of engineering theory…." (p.314)

Accordingly, homeostatic control systems have a number of "required features": 1) a set point which defines the target output of the system; 2) feedback; 3) precision in resetting the output back to the set point, following a perturbation; and (normally) 4) sensors which measure the difference between the actual output and the set point (p.309).

Thus control theory offers neuroscientists clear and experimentally testable criteria for determining whether a system undergoes homeostatic regulation, by looking for these required features (e.g. the existence of a set point) in a system. The operating conditions of homeostatic regulation, and the biophysical mechanisms of feedback, sensors, etc., are also open to experimental investigation. Reported examples of properties under homeostatic control are muscle excitation at the neuromuscular junction (p.309) and bursting properties of invertebrate neurons (p.311). More recently, O'Leary et al. (2014, p.818) argue that ion channel expression in their simplified model of invertebrate neurons can be understood as an implementation of *integral control*, a standard control-theoretic architecture.

Figure 2 (if space) schematic for integral control

3. CRASH TESTING THE FRAMEWORK

Before considering the question of whether the engineering framework becomes structurally unsound when applied to some kinds of neural systems, I would like to draw our attention to some of its features. The basic ideas are clearly illustrated in Sterling and Freed's (2007) example of the bridge. When one considers the robustness of an engineered artefact like the bridge, it is presupposed that the system is built up from component parts in such a way as to achieve a specific function. The robustness of the bridge is conceptually distinct from its other designed features or functions, and it can trade off against some of them. For example, the more robust the bridge is to the passage of the occasional heavy vehicle, the more expensive it will be to build (because requiring more steel) (p.563). Moreover, the perturbations against which the system is robust are thought of as atypical events, also conceptually distinct from the normal operations of the system.

There is also the tendency to think of robustness as allowing the system, following a perturbation, to return to its initial stable state. Some experiments specifically involve the operationalization of the robustness of a system as the reversion to a prior state. For example, reporting on an experiment in which mouse premotor cortex in one hemisphere was inhibited using optogenetics during the preparation period for the animal's movement, Svoboda (2015)[3] writes, that "[t]his preparatory activity is remarkably robust to large-scale unilateral optogenetic perturbations: detailed dynamics that drive specific future movements are quickly and selectively restored by the network." This notion of robustness as the ability of the system to revert to a

[3] To my knowledge, these results have not yet been published in a journal. I have contacted the author to find out if the study is under review or in press.

prior functional states is similar to the idea of *homeostasis* as the ability of a system to stabilize some quantity in spite of external changes.

Figure 3 (if space) After Kitano (2004, Figure 1)

Eve Marder's laboratory has carried out a long term investigation into the ability of neurons to maintain stable electrophysiological properties despite continual turnover of the ion channels embedded in the cell membrane which are responsible for its electrical excitability. This research project is one of the central examples of the study of robustness in neural systems. Marder and her collaborators make ample use of the engineering framework when reviewing other results and reporting their findings. For example, O'Leary et al. (2013, p.E2645) write:

> "Both theoretical and experimental studies suggest that maintaining stable intrinsic excitability is accomplished via homeostatic, negative feedback processes that use intracellular Ca^{2+} concentrations as a sensor of activity and then alter[s] the synthesis, insertion, and degradation of membrane conductances to achieve a target activity level."

What is striking about the characterization of electrophysiological stability in the face of ion channel turnover as a kind of robustness in the face of a perturbation (e.g. p.E2651), is the fact that the turnover is just part of the normal physiology of the cell. There is no functional and stable state of the cell in which this turnover does not occur—a fact which these authors also highlight.[4] This brings our attention to some strains in the application of the engineering framework to this biological system.

[4] "neurons in the brains of long-lived animals must maintain reliable function over the animal's lifetime while all of their ion channels and receptors are replaced in the membrane over hours, days, or weeks. Consequently, ongoing turnover of ion channels of various types must occur without compromising

In the basic engineering characterisation of robustness, sketched above, perturbations are different from the normal circumstances in which the system is expected to operate. "Perturbation" carries the everyday connotation of an event which throws the system off balance and is deleterious to its normal functioning. We cannot think of the events of ion channel turnover as perturbations in this sense; they are business as usual for the cell.

Furthermore, it is not in the nature of the system to seek to return to a prior, stable arrangement of its parts. A crucial property of the nervous system is its plasticity: the tendency for its component parts and the connections linking them to be continually sculpted by experience. The homeostatic mechanisms which Marder and colleagues investigate need to be understood as maintaining specific properties (such as a cell's Ca^{2+} concentration) at a certain point, but not (nor do these researchers claim it) some generalised operation for achieving system-wide internal stability (see §4.4).

In the basic engineering conception of robustness, there is a clear conceptual distinction between the features of a system which allow it to carry out its intended function, and those which make the system robust (even if in reality one individual feature can serve both purposes). In the case of the neuron which has continual ion channel turnover and no definite stable state to return to following these "perturbations", it is not clear that we can we make this distinction. A more natural way to think about this and other biological systems is as ones, unlike engineered artefacts, "designed" to keep changing

---

the essential excitability properties of the neuron" (O'Leary et al. 2013, p.E2645).

and "designed" to maintain functional stability in the midst of this constant change.[5]

The tensions and strains associated with the application of the basic engineering framework to biological systems can be felt more sharply if we appeal to a process metaphysics of biological "things" (Dupré 2012). According to this view, organisms are not substances but *processes*—items whose existence depends on the taking place of certain changes. This highlights the fact that the life of organisms depends on a continual turnover of its component parts, and that the system as a whole, while living, persists longer than its parts. Yet features and functions of the organism remain relatively stable. For example, memories can endure for decades even though the neurons that form them have undergone material change. This stability must be achieved—somehow. And so processes for robustness are not cleanly distinct from the general maintenance processes which keep the organism alive.

The processual nature of neurons is nicely described by Marder and Goaillard (2012, p.563):

> "each neuron is constantly rebuilding itself from its constituent proteins, using all of the molecular and biochemical machinery of the cell."
> (and see F n 4)

[5] This blurring of the lines between mechanisms for robustness and mechanisms for life is highlighted by Edelman & Gally (2001: 13763) in their discussion of the difference between redundancy and degeneracy in biological systems: "the term redundancy somewhat misleadingly suggests a property selected exclusively during evolution, either for excess capacity or for fail-safe security. We take the contrary position that degeneracy is not a property simply selected by evolution, but rather is a prerequisite for and an inescapable product of the process of natural selection itself." They also discuss another disanalogy between engineered and biological systems—the applicability of "design" talk.

We can contrast this with the substance metaphysics that we usually assume when thinking about engineered artefacts. A bridge or an aeroplane is what it is because of the parts which comprise it. Its existence does not depend on the occurrence of any process. This is not to deny that an expert in the theory of matter might well argue that the steel of the bridge maintains its integrity because of some fundamental processes. The point is that when characterising the robustness of the bridge or the aeroplane we would not resort to such sophistication. Rather, we think of the bridge as a substance and not a process—a steel structure which, in order to maintain its function in the face of perturbation, must resist rather than effect the swapping around of its component parts.

4. EXAMINING REASONS TO RE-ENGINEER

Now that we have noted these disanalogies between biological organisms and engineered things, we ought to worry that the framework borrowed from engineering is misleading when thinking about robustness in the brain and other biological systems. *Is it time to re-engineer our conceptual tools for thinking about robustness to make them more suitable for characterising living things?* In this section I consider four possible answers to this question.

4.1 *No. The terms in the engineering framework are just words that are used to facilitate communication of the neuroscientific results.*[6]

One potential response to the concerns raised in the previous section is that they stem from a superficial fixation on the vocabulary neuroscientists use when writing about their research. Just because the authors discussed above

[6] A response along these lines was suggested to me by Timothy O'Leary, in conversation.

have employed certain words first introduced by engineers, it does not follow that their understanding of neurophysiology is distorted by comparisons with engineering. For example, I mentioned that the word "perturbation" has a negative connotation which makes it seem inappropriate when describing non-pathological and frequent events like ion channel turnover. It could well be that in the context of this research the term takes on a different meaning—for example, as any event that the system cannot directly control,[7] such as changes in protein configuration due to thermal noise.

I believe that this response is warranted by what we know of the methodology of some of the investigations discussed above, but not all of them. In the case of Sterling and Freed (2007) I was careful to show that the engineering conceptions directly shaped how these neuroscientists operationalized and quantified robustness, and how they identified mechanisms by which robustness is achieved. There is no indication that they used terms such as "safety factor" to mean something radically different in the context of neuroscience.

A very explicit statement of the aim to apply engineering principles directly to the understanding of the premotor cortex comes from Svoboda (2015):

> "preparatory activity is distributed in a redundant manner across weakly coupled modules. These are the same principles used to build robustness into engineered control systems. Our studies therefore provide an example of consilience between neuroscience and engineering."

Thus the convergence between a neurophysiological and the engineering perspective on the mouse motor planning system is taken to be an important result of this study. This echoes Sterling and Laughlin's (2015, pp. xiii-xv) proposal that enquiring to see how engineering principles are implemented in

[7] I thank Timothy O'Leary for this suggestion.

neural systems, and the attempt thereby to reverse-engineer the brain, leads to insights not otherwise available through routine data collection.

4.2 *No. The inadequacies you point out with the engineering framework are based on a caricature of mechanical engineering, not the actual complex discipline.*[8]

My characterisation of the engineering framework assumes that mechanical engineering (the design of bridges, aeroplanes and such like) is paradigmatic of the engineering approach in general. But of course there are many different kinds of engineering, from mechanical to electronic to communications and chemical. It could well be that the mismatch between understanding the robustness of a highly dynamic entity like the brain, and the rather static conception of robust objects that falls out of the basic engineering framework is just an artefact of only focussing narrowly on the kind of engineering that is actually furthest away from neuroscience.

It would take me beyond the scope of this short article (and well beyond my own knowledge of the subject) to sketch out the various possible frameworks associated with each field of engineering specifically, and to see which conception of robustness is most suitable for biology. However, what I will say is that there is evidence in the studies discussed above that neuroscientists themselves do sometimes draw from the mechanically based caricature. This is particularly true of Sterling and Freed (2007). In contrast, when Davis (2006) and O'Leary (2014) make direct appeal to engineering they refer specifically to models in control theory.[9] This invites questions, still, about whether the paradigm examples of controlled systems (e.g. a car driven on

[8] This concern was raised by Arnon Levy and Timothy O'Leary.
[9] See also Zhang and Chase (2015) on the physical control system perspective on brain computer interfaces for motor rehabilitation.

cruise control, a Watt governor, or an aeroplane flown on autopilot) are dynamical enough capture the processual nature of the nervous system.

4.3 *Yes. The brain is so different from an engineered artefact that the framework is misleading and inappropriate.*

In Sections 4.1 and 4.2 I discussed two reasons for thinking that we should not be concerned about any radical disanalogy between robustness in biological and engineered systems. While I agree that these are important points to keep in mind, I do not think that they diffuse the fundamental concern that when neuroscientists borrow engineers' terms in order to study robustness, they risk mischaracterising the brain as more like an engineered artefact than it actually is. Is the appropriate conclusion, then, that a neural circuit is so different from a bridge or an aeroplane that the engineering framework is simply misleading and should be discarded?

The best way to make this strong negative case is to consider some historical examples in which reasoning by analogy with engineered systems seems to have lead neuroscientists and theorists astray. One example comes from von Békésy, a physicist and communications engineer who turned his attention to inhibition in the nervous system. In his book *Sensory Inhibition* he notes that there are feedback loops everywhere in nervous system and he asks how it is that system manages to avoid ending up in a dysfunctional oscillatory state (1967, p.25). It seems that von Békésy is importing his understanding of systems containing feedback from engineering, and in that context oscillations are normally problematic and efforts must be made to dampen them. These days neuroscientists seek to understand how oscillations in the healthy brain (i.e. its characteristic patterns of endogenous activity) are actually responsible for cognitive functions, and how these oscillations differ

from the ones associated with pathologies such as epilepsy and Parkinson's disease.[10]

Another example is the comparison of the effects of "noise" in brains and artificial signalling systems…… GET EXAMPLE
This is very different from how neuroscientists understand noise today, which begins with the idea that brains evolved under constraints imposed by noisy "components", which has therefore shaped all aspects of neural computation (Faisal et al. 2008). It would be a mistake to think of the brain processing information in the same way as an electronic computer, but with added redundancy to offset the noisiness of individual processing streams.

The cautionary tales just told give some concrete indications of how imposition of the engineering framework on to neural systems can lead to conclusions which in retrospect appear false and misguided. But it would be too hasty infer from these two examples that current work on robustness in neuroscience is of dubious standing whenever it appeals to the concepts of engineering. A more general argument is the following: *the brain is not like a bridge (or a computer, or an aeroplane on autopilot….); therefore whenever neuroscientists appeal to terms borrowed from the analysis of such systems, they risk saying things that are simply false because they fail to notice relevant disanalogies.* This lays all the sceptical cards on the table. In the last part of the paper I attempt to mitigate these worries.

[10] For a scientific overview see Buzsáki (2006). For discussion of philosophical implications, see Bechtel and Abrahamsen (2013). See also Knuuttila and Loettgers (2013, p.160) on a parallel difference across engineering and cell biology, where oscillations are found to have a functional role.

4.4 *No. Use of the engineering framework should be thought of as a simplifying strategy.*

Neuroscientist Steven Rose (2012:61) writes that:

> "one of the most common but misleading terms in the biology student's lexicon is homeostasis….[the] concept of the stability of the body's internal environment. But such stability is achieved by dynamic responses; stasis is death, and homeodynamics needs to replace homeostasis as the relevant concept"[11]

This seems to capture the problem that was first noted in Section 3, that we should not be mislead by the engineering framework into thinking of neural systems as seeking to maintain an initial stable state. But we also noted that the neuroscientists employing control-theoretic models of homeostatic mechanisms are not thinking of their systems as seeking stability in this very general way. Instead, they are modelling the stability of a specific variable—in the case of O'Leary et al. (2014), the concentration of Ca^{2+}—and investigating the mechanisms by which it is controlled. To this end, it is reasonable to interpret the system as an integral controller (p.818).[12] Thus it is still useful to talk about homeostasis with respect to Ca^{2+} concentration, even while thinking of the system as a whole, and in reality, as a "homeodynamic" one.

---

[11] Compare Sterling (2012) on the concept of *allostasis* – stability through change with an emphasis on predictive regulation. Day (2005) and O'Leary and Wyllie (2011), in contrast, argue that the concept of homeostasis easily accommodates these dynamic and predictive aspects, and that the term *allostasis* is therefore superfluous. It is an interesting question (but beyond the scope of this paper) whether the narrow or wide definition of *homeostasis* is currently more prevalent amongst biologists and neuroscientists.

[12] Note that O'Leary et al. (2014) study of homeostasis is via a *model* of a neuron. But the model is realistic enough that it is expected to shed light on actual biophysical mechanisms.

I think of neuroscientists whose investigation of robustness in the brain is scaffolded by the engineering framework as providing *idealized mechanistic explanations.* Their explanatory target is, for example, the process by which overall neuronal activity level is controlled via regulation of ion channel gene transcription through a Ca^{2+} sensitive feedback loop. This is standard fodder for mechanistic explanation. At the same time, the framework of engineering—in this case the schematic of the integral controller—serves to direct attention to specific parts and processes in the extremely complex cellular machinery and to interpret them in control theoretic terms (sensors, feedback loops, etc.), while bracketing other aspects not immediately relevant to the explanation of robustness.

Bechtel (2015, p.92) has presented the case that:

> "mechanisms are [to be] viewed not as entities in the world, but as posits in mechanistic explanations that provide idealized accounts of what is in the world."

His example is the idealization (understood as "falsehood") that scientists introduce by putting boundaries around putative mechanisms which in nature do not exist. In the cases explored in this paper, the idealization comes in through the analogical reasoning of treating a neuronal system *as if* it is an engineered artefact. This, like the positing of boundaries, is a useful way to simplify the explanandum. It enables neuroscientists to bracket some of the known facts about the brain's messy, Heraclitean nature. But it means, perhaps, that there is a stark difference between the brain viewed *sub specie aeternatis* (what some neuroscientists call the "ground truth" of the brain) and viewed *sub specie mechinae* (in the guise of a machine).

ACKNOWLEDGEMENTS

I am greatly indebted to Timothy O'Leary, Nancy Nersessian and Peter Sterling for their feedback on this work. I would also like to thank the

participants of the Fall 2015 workshop on Robustness in Neuroscience for discussion of the ideas behind this paper, and the audience at the Spring 2016 Re-Engineering Biology conference for their questions and comments on it. Both of these events were hosted by the Philosophy of Science Center at the University of Pittsburgh.

REFERENCES

Bechtel, W. and Abrahamsen, A. (2013). Thinking dynamically about biological mechanisms: Networks of coupled oscillators. *Foundations of Science*, 18:707–723

Bechtel, W. (2015). Can mechanistic explanation be reconciled with scale-free constitution and dynamics? *Studies in History and Philosophy of Science Part C,* 53: 84–93.

von Békésy, G. (1967). *Sensory Inhibition.* Princeton, NJ: Princeton University Press.

Buzsáki, G. (2006). *Rhythms of the Brain.* Oxford: Oxford University Press

Davis, G.W. (2006). Homeostatic control of neural activity: from phenomenology to molecular design. Annu. Rev. Neurosci. *29*, 307–323

Day TA (2005). Defining stress as a prelude to mapping its neurocircuitry: no help from allostasis. *Prog Neuropsychopharmacol Biol Psychiatry* **29**, 1195–1200

Dupré, J. (2012) *Processes of Life.* Oxford: Oxford University Press

Edelman GM, Gally JA (2001) Degeneracy and complexity in biological systems. Proc Natl Acad Sci USA 98(24):13763–13768.

Faisal, A., L. P. J. Selen and D. M. Wolpert (2008) Noise in the Nervous System. *Nature Reviews Neuroscience.* 9:292-303.

Kitano, H. (2004) Biological robustness. *Nature Reviews Genetics.* 5: 826-837.

Knuuttila, T. and A. Loettgers (2013). Basic science through engineering? Synthetic modeling and the idea of biology-inspired engineering. *Studies in History and Philosophy of Science, Part C 48*, 158–169.

Marder, E. and Goaillard, J.-M. (2012) Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*. 7:563-574

von Neumann, J. (2000). *The Computer and the Brain*. New Haven: Yale University Press.

O'Leary T, Williams AH, Caplan JC, Marder E (2013) Correlations in ion channel expression emerge from homeostatic tuning rules. *PNAS.* 110(28): 809–821

O'Leary T, Williams AH, Franci A, Marder E (2014) Cell types, network homeostasis and pathological compensation from a biologically plausible ion channel expression model. *Neuron* 82(4): E2645–E2654.

O'Leary, T. and D. J. A. Wyllie (2011) Neuronal homeostasis: time for a change? *J Physiol* 589.20:4811–4826

Rose, S. (2012). The need for a critical neuroscience. In *Critical Neuroscience: A Handbook of the Social and Cultural Contexts of Neuroscience,* S. Choudhury and J. Slaby (eds.) Hoboken, NJ: Wiley-Blackwell.

Sporns, O. (2010). *Networks of the Brain.* Cambridge, MA: MIT Press.

Sterling, P. and M. Freed (2007). How robust is a neural circuit? *Visual Neuroscience*, **24**, 563–571.

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology & Behavior* 106:5–15

Sterling, P. and S. B. Laughlin (2015). *Principles of Neural Design*. Cambridge, MA: MIT Press.

Svoboda, K. (2015). Probing Frontal Cortical Networks during Motor Planning. Abstract, Center for the Neural Basis of Cognition, 10 November 2015. http://www.braininstitute.pitt.edu/event/probing-frontal-cortical-networks-during-motor-planning

Zhang, Y. and S. M. Chase (2015). Recasting brain-machine interface design from a physical control system perspective. *J Comput Neurosci* 39:107–118

PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association -160-

Eight Myths about Scientific Realism

ABSTRACT: Selective realist projects have made significant improvements over the last two decades. Judging by the literature, however, antirealist quarters seem little impressed with the results. Section I considers the selectivist case and its perceived shortcomings. One shortcoming is that selectivist offerings are nuanced in ways that deprive them of features that—according to many—cannot be absent from any realism "worth having". Section II (the main part of the paper) considers eight features widely required of realist positions, none of them honored by selectivist projects. Modulo those requirements, even if selectivists managed to clear other shortcomings of their project selectivism would still not be a position worth considering. Next the historical background and present credentials of the requirements in question are examined. All are found to rest on myths and confusions about science and knowledge. If this is correct, realists and antirealists should reject the requirements.

I. Background

The antirealist waves of the 1980s stifled naïve realist projects, but they also gave rise to critical realist reactions, particularly a shift in the way theories are accepted at face value from whole constructs to selected "theory-parts" (existence claims, narratives and structures regarding features beyond the reach of unaided perception). Moves in this "selectivist" direction were variously developed in the 1980s and 1990s, most influentially by Leplin (1984), Worrall (1989b), Kitcher (1993), Leplin (1997), and Psillos (1999). Selectivists see in the history of science a past littered not just with failures but also clear successes, especially after the consolidation of methodologies focused on impressive novel prediction in the early 19th century. The successes selectivists point to involve law-like structures all over physics, functional (as opposed to formally "fundamental") entities like the particles invoked by the kinetic theory of matter, numerous extinct species hypothesized by Darwin and his circle, structures and processes from microbiology, much in Mendelian genetics, myriads of molecular structures, and most of the subatomic entities deemed well-established since the 1950s, along countless causal networks, histories and functional entities in virtually all theories with warrant in terms of impressive novel predictive success. Selectivists thus respond to skeptical readings of the history of science with optimistic readings, which they argue are better justified than Laudan (1981)'s skeptical appeals. History, Leplin (1984) noted early in the debate on selectivism, is not opposed to realism any more than our experience of ordinary objects is unambiguously veridical.

In selectivist terms, successful scientific theories may provide imperfect representations of unobservable aspects of some of their intended domains, but they do get those aspects right to some significant extent—and *that is what matters to a realist stance*. Realism has to do with

having warranted augmentative inference at levels that reach into unobservables, i.e. beyond the level allowed by its contrast position—constructive empiricism.

Developing selectivism into a mature project has not proved easy. The initial criteria proposed for identifying theory-parts worthy of realist commitment were either too vague or picked up through "retrospective" projection of current. As Kyle Stanford (2006) cautioned, mere retrospective projection of current science reflects limitations of human imagination as easily as it does truth-content and can be variously misleading; also, it can be self-serving, and worse still it severely weakens selectivism by giving up the traditional realist goal of identifying the truthful parts of a theory while the theory is still alive. Realists need to develop compelling criteria for prospective projection, applicable to theories in full flight, and over the last decade selectivists have moved imaginatively to respond to this challenge. One promising contribution is a stronger emphasis on impressive novel predictions as a marker of success and truth content. This trend is multiply developed in works that revisit in detail the cases most used by antirealists as springboard exemplars of gross epistemic failure, as well as studies of other seemingly germane cases from the last 200 years (e.g. Saatsi 2005, Saatsi & Vickers 2011, Votsis 2011, Vickers 2013, @@@). While the debate is far from over, upgraded proposals are on view in the selectivist analyses just cited. At the very least, the initial antirealist arguments from radical underdetermination and so-called "skeptical inductions" have been weakened by selectivist challenges to the antirealist arguments at work. Still, many critics join Stanford in thinking that selectivism lacks a convincing realist criterion for prospective identification of theory-parts. As said, promising selectivist developments seem on view in this regard, but there is something else.

Something seems to be making the selectivist project intellectually unattractive in some quarters, independently of the issue about the criterion for theory-parts. There is, in particular, a perception (not least among many sympathizers of realism) that selectivism advances its case at the cost of diluting its realist import, resulting in *a stance "not worth having*." By the lights of selective realism, an empirically successful theory T contributes significant truths about unobservables but

(1) typically, what makes T approximately true is that abstract versions of some of its parts are truthful, making the realist stance applicable to selected fragments of T rather than the integral whole initially intended;

(2) such truth as T contains need not have universal applicability;

(3) T need not offer literal truth at its most fundamental level;
(4) the significance of T's central terms is high in unificationist rather than epistemic terms;
(5) T adds significantly to our knowledge of unobservables in the intended domain, but there is no reason to expect T to be "right for the most part" at *any* level (what matters is that it yields epistemic gain at theoretical levels).
(6) T may not instantiate uniformly convergent progress towards any "final description;"
(7) the intelligibility T confers to its intended domain is generally incomplete.

Each of the above tenets clashes head on with widespread assumptions and expectations regarding a realist stance about theories. The latter, many believe, *should* (1) constitute integral wholes, (2) apply universally, (3) give correct theoretical description, (4) have central terms that refer, (5) be, at least, *right for the most part.* (6) display epistemic progress, and (7) offer substantial intelligibility of the intended domains, Behind these expectations about scientific theories and what theoretical claims amount to is a view on what a *realist position worth having* comprises: to be worth having, a realist position must encompass strong versions of most of the listed assumptions. Antirealists (and not a few realists) routinely take these assumptions for granted. This aspect of the debate needs discussion because, as noted, the assumptions in question are clearly at odds with the selectivist strategy, which—generalizing Worrall (2016) a bit—might be *the only viable realist game in town.*

II. Taxing Assumptions

There is a view, shared by numerous scientists, according to which scientific realism cannot be a position worth having unless it encompasses most of the traits listed at the end of the last section. One problem with those traits is that they provide antirealists with fodder for criticizing positions that embrace them and realists for dismissing positions that lack them. Let us consider the listed items in detail.

(1) **Theories as Integral Wholes**. Selectivism rejects the view that theories and conceptual networks are intellectual constructs made of non-separable parts. The integral wholes vision commits realism to nothing less than complete theories. Motivations for it come from at least two

fronts. One includes linguistic holism and/or the statement view of theories, endorsed in the 1960s and 1970s by thinkers as superficially different as Ernest Nagel and Thomas Kuhn. Another motivation, good for a weaker version of the vision, has been the presumption that some concepts are grounded in "metaphysical necessities," a position widely held in natural science until the early 1900s. In the 19th century it was thought that breaking of a theory into independently assertible parts had drastic limits. A case in point was the need felt for positing an ether of light, as at the time waves were conceived of within a traditional metaphysics that regarded them as propagating disturbances and thus as ontologically dependent entities that *required* the existence of something being disturbed (@@@). Institutional deference towards similarly presumed conceptual necessities is massively lower now. One major inflection point was the acceptance of Einstein's Special Relativity, which opened the road to changes in both the conception of light and the requirements of intelligibility in physics.

Nobody thinks now that light is completely as Fresnel or Maxwell imagined, yet— having no conceptual links closed to the possibility of scientific revision— there is little question that Fresnel's theory got many things right, e.g. what might be termed "Fresnel's Core": light is made of microscopic transversal undulations, and these undulations follow the Fresnel laws of reflection and refraction. Abstracted from reference to the wave substratum, this schematic part of the theory spells out a descriptive core that all subsequent theories of light have retained. Once conceptual networks are recognized as relations sustained by revisable inductive conjunctions, scientific "good sense" allows shifts in science towards theory-parts cut out from the rest. There is a historical supplement to this. There has never been much serious allegiance to theory "unbreakability" *in scientific practice*. As scientists developed their ideas, virtually all took a realist stance towards just selected parts of a theory at hand while taking a non-realist stance towards other parts (e.g. Newton's approach towards Kepler's cosmology and Galileo's mechanics; 19th century wave theorists towards particle theories of light, Einstein towards Fresnel's Core, Einstein towards Newtonian mechanics, molecular geneticists towards Mendelian genetics, and so forth). Being selective about what to take at face value in a theory is exactly what selective realists do, also what we all do in ordinary life. The idea that proper theories are unbreakable integral wholes just rests on myth.

(2) **Universality**. Another widespread assumption is that, for realism, proper scientific theories must hold universally. We find this view expressed in e.g. van Fraassen (1980: 86): from a realist perspective, he claims, ''a theory cannot be true unless it can be *extended* consistently, without correction, to all of nature''

This request rests on myth. There is no reason to think that interesting theories can be so extended even at the lowest phenomenal level. Generalizations limited to the observable level typically turn out to be true only over restricted ranges, just as with theoretical generalizations. The standards of acceptability should not be arbitrarily raised against scientific theories. So, past successful theories could not be extended consistently, without correction, to all of nature. However, as selectivists show, those theories made significant cognitive gains at significant levels, where various assortments of the theoretical descriptions they licensed remain both accurate and illuminating. The universality objection, it seems, burdens realism with a suicidal demand.

(3) **Truthful description**. Realists are allegedly claim that what a theory T says about entities, properties, relations and processes should be construed literally; and to take a realist stance towards T is to believe that what it says is literally true. This view comprises three major lines: (3a) literalism, (3b) accuracy realism, and (3c) a methodological supplement.

(3a) Like their biblical counterparts, theory-literalists think one mistake in a narrative is one mistake too many. Phlogiston theory got some of its central claims wrong, as did also Fresnel's theory, Mendel theory, Bohr's 1913 theory of the hydrogen atom, and countless other theories, so those theories were all completely wrong.

The antirealist uses of literalism are straightforward. If departures from literal accuracy, however small, make theories count as different, then the chances of a scientist ever picking *the* right theory will be wretchedly small (argument of the bad lots). And the probability of conjecturing the one (and only one) truthful theory will be hopelessly small (problem of the base rate). And, so, at any given time, the chances that the one truthful theory is among the as yet "conceived alternatives" will be overwhelmingly low.

Happily for realists, the expectations in (3a) belong in fairy-tales. Scientific theorizing is rarely strictly literalist. Scientists effectively abandoned literalism early in modern times, as they

began to articulate explanatory idealizations that carried an expectation that nothing in nature exactly realized them. For example, the aim of the kinetic theory of matter developed around 1860 was to causally account for approximate empirical laws that had been gathered in the two previous centuries about the macroscopic behavior of gasses (e.g. PV = nRT) and materials (e.g. thermal expansion). Crucially, in the case of gases, the accounts invoked structureless point-particles—the so-called "ideal gas"—that the theorists involved did not believe existed in nature. The ideal gas was *explicitly* an idealization, with a two-fold expectation at work: (i) actual gasses are made of non-ideal corpuscles moving at random and located at relatively large distances from one another "on average"; and (ii) the behavior of those actual corpuscles *instantiated that of the ideal gas to a significant degree* within a certain restricted domain. There was no question that ideal gasses literally construed had to be "real" in order to take the theory realistically. Scientific theories are likewise *generally* false in strictly literal fashion. As with maps, the point of realist interest is the extent to which a theory's depictions match the *intended* domain. Theoretical representations of empirical domains resemble maps far more than they do assertions (e.g. Giere 2006). Selectivists proceed accordingly: taking a realist stance towards a theory T amounts to claiming only that some of the explanations and descriptions distinct to T are correct by *acceptable standards*.

(3b) In mathematized disciplines literalism easily ups its ante. According to a long lived assumption of quantitative exactitude, there are in nature quantities of which concrete systems have definite values, and in a correct theory the claims it makes correspond to the world with total accuracy. This ideal is found in early modern scientists, notably theorists with strong Platonist leanings such as Kepler.

Dear though these expectations of divine accuracy and depth are, they rest on myth. Such correspondence as mathematized theories have to the world is not conditioned to radical accuracy. As Bertrand Russell noted on behalf of sound epistemology,

> "Although this may seem a paradox, all exact science is dominated by the idea of approximation. When a man tells you that he knows the exact truth about anything, you are safe in inferring that he is an inexact man. Every careful measurement in science is

> always given with the probable error [...] every observer admits that he is likely wrong, and knows about how much wrong he is likely to be." (1931: 42)

More recently, in a more comprehensive vein, Paul Teller (2015) complains that "accuracy realism" assumes that the quantities invoked by a theory actually refer. But—he notes—this misunderstands the fabric of theoretical representation, because theories generally formulate *idealizations* that burden quantitative attributions with failure of specificity in picking concrete cases. In the narrowest literal sense, the claim "the meter-standard kept in Paris is 1 meter long" may be true only by *definition*—any attempt to check it with absolute precision against any external objective length would be frustrated by, to begin with, ineliminable thermal and quantum mechanical fluctuations. The point is that one-to-one matching makes no sense as a goal in scientific language, given that so many descriptive words in science are intrinsically vague and/or refer to idealizations. Actual reference to lengths presumes just perspectivally *acceptable* (never absolute) accuracy. At the lowest empirical levels also, completely exact assertions are generally neither relevant nor true. This connects with a related point, namely, the *irrelevancy* of these literalist and accuracy assumptions to the actual realism/antirealism debate. Shaped by the discussions started in the 1980s, the dispute is now primarily about whether or not warranted augmentative scientific inferences reach into unobservable domains. Ordinary realism about chairs, cats and mountains fails the ideals of radical literalism and accuracy no less than scientific realism.

(3c) The methodological supplement claims that science would be merely an instrumentalist affair unless theorists aim to produce a complete description of the way things are, with scientists as pursuers of God-like reportage (perfect "mirror reflection"): scientific theories advance towards the truth, all the truth, and nothing but the truth (see e.g. Sankey 2008's discussion of this). Although this position lost much of its ancient appeal in the 18th century, to this day some top theoreticians continue to wax lyrical expressing it, especially in "editorials".

> "The 'theory of everything' is one of the most cherished dreams of science. If it is ever discovered, it will describe the workings of the universe at the most fundamental level and thus encompass our entire understanding of nature. It would also answer such

> enduring puzzles as what dark matter is, the reason time flows in only one direction and how gravity works. Small wonder that Stephen Hawking famously said that such a theory would be 'the ultimate triumph of human reason – for then we should know the mind of God' ". (New Scientists, 4 March 2010[1])

This colorful supplement lacks warrant if, as selectivists claim, the realist stance can be consistently and fruitfully applied to selected theory-parts.

The realist badge of honor is not awarded for telling the truth, all the truth, and nothing like the truth about anything—let alone reading the mind of God. It is a distinction for *finite cognitive achievements forged with crooked tools*. See also (6) below.

(4). **Realist Significance of the "Central Tenets" of a Theory**. A related common assumption is this: Even if truthful description may have limits, taking a theory T realistically requires commitment to T's central tenets (i.e. those about the entities, principles and laws that individuate T). In Laudan's version, realism about T commits to the view that the T's central terms *successfully refer*.

There is little question that in numerous scientific theories the central terms fail to refer—on this point we all have a debt of gratitude to Laudan. However, once theories are no longer approached as unbreakable wholes the emphasis on central terms wanes. If anything, the reference that matters is that of theory-parts. Then, on the explanatory side, the scientific focus is on the structures of possibilities of its intended domain D. As such, a theory is not exclusively about the entities and relations invoked at the level of its central terms. Primarily the theory is about D, whose relevant entities and structures include those that may be found at intermediate levels of description—like Fresnel's Core. A theory may thus be individuated by its central tenets, but the latter do not exhaust the theory's realist import. The appropriate realist focus is those theoretical claims derivable from the theory and for which there is strong evidence (and so a strong expectation of truthfulness), not whether the terms involved are "central", "intermediate", or "peripheral".

---

[1] "Knowing the mind of God: Seven theories of everything", New Scientists, 4 March 2010: https://www.newscientist.com/article/dn18612-knowing-the-mind-of-god-seven-theories-of-everything/

(5). **Being "right for the most part".** Another related assumption links the realist stance towards a theory with the claim that the theory is right "for the most part". Michael Devitt, for example, voices this assumption when he defines scientific realism as the doctrine according to which "Most of the essential unobservables of well-established current scientific theories exist mind-independently and mostly have the properties attributed to them by science" (2005: 769). In his view, theories that are well-established theories *by today's* methodological standards are right *for the most part*.

This supposition sounds reasonable at first hearing but it too seems suicidal for realism. Virtually all the past theories realists want to be realist about seem to have turned out to be wrong "for the most part"—unless "being right" is granted with postmodern largesse. Newtonian mechanics is "right" for a comparatively tiny regime of speeds and fields. Bohr's theory of the atom gets impressive aspects right but otherwise is wrong for the most part of the entire quantum domain. Mendel's theory invites a similar reaction. For all we know, our excellent present physics may be wrong for most of the *total universe*. So, scientifically successful theories seem "wrong for the most part". But they have great realist import, nonetheless. That import comes from the fact that they get right novel *significant unobservable* aspects of their intended domains. As David Bohm urged long ago, piecemeal caution needs to be exercised in one's realist commitment to the entities, regularities and processes invoked by well-established current scientific theories (1957, Chapter V). Two lines of reasoning in particular support this prudence (@@@): (1) Qualities, properties of matter, and categories of laws expressed in terms of some finite set of qualities and laws are generally applicable only within limited contexts (in terms of ranges of conditions and degrees of approximation). (2) There is no reason to suppose that new qualities and laws will *always* lead to mere correction refinements that converge in some simple and uniform way. This may occur in some contexts and within some definite range of conditions, but in different contexts and under changed conditions the qualities, properties and laws may be quite novel and lead to dramatic effects relative to what previous theorizing would have led to expect. For example, for bodies moving with speeds negligible compared to the speed of light, the laws of relativity lead to small corrections of the laws of Newtonian mechanics. But they also lead to such qualitatively new results as the "rest energy" of matter. Further laws yet to discover may be vastly more bizarre.

(6) **Progress**: The realist expectation that successful science achieves cumulative truth content about unobservables is frequently nailed to the idea that "modern science is converging on a single picture of the world". Claims along these lines come in several flavors, in particular (a) linear epistemic progressivism and (b) "metaphysical" realism.

(6a) Convergent progress. Léo Errera expressed the idea in his *Botanique Générale* of 1908: "Truth is on a curve whose asymptote our spirit follows eternally[2]." This expectation has recurrent mystical roots in science. John Herschel, for example, is cited by Marcel de Serres as saying "All human discoveries seem to be made only for the purpose of confirming more strongly the truths come from on high, and contained in the sacred writings[3]."

Convergent progressivism runs against a recurrent realization in modern science. As selectivists recognize, successful theories give knowledge but they usually err at numerous levels of description. Successful theories don't give us everything there is to know about any intended domain, let alone 'The World.' Finite sets of simple laws can provide correct descriptions and predictions when we constrain their context enough, notes Bohm (1957), but we should expect unrestricted theories to be false. Many defenders of scientific objectivity have followed suit, stressing the shift from traditional searches for a comprehensive world-view to explicitly perspectival searches for piece-meal knowledge about domains of current scientific interest, leading to assertions of corresponding partiality.

(6b) In no better shape is the claim that realism is committed to the existence of one true and complete description of the world, whose truth bears one-to-one correspondence to 'mind-independent reality, so that the purpose of science is to discover that description. Critics persuasively dismiss this brand or realism. But no knowledgeable realist has held such a position in generations. It is a thesis recalled from the grave in the late 1970s and 1980s by Hilary Putnam under the label "metaphysical realism," a view he presented as an example of a hopelessly jumbled project (e.g. Putnam, 1978: 49, and 1990: Preface).

---

[2] *Recueil d'Œuvres de Léo Errera: Botanique Générale* (1908), 193. As translated in John Arthur Thomson, *Introduction to Science* (1911): 57

[3] Marcel de Serres, 1845. "On the Physical Facts in the Bible Compared with the Discoveries of the Modern Sciences". *The Edinburgh New Philosophical Journal* (Vol. 38): 260. [239-271]

(7) **Intelligibility**: Another claim often associated with realism is that science aims to provide truthful explanations that make the phenomena at hand intelligible. This condition comes in (a) radical and (b) moderate strengths. The radical version calls for explanations that leave the intellect content and with no further whys. The weak condition calls for explanations that make the target phenomena *more* but not necessarily fully intelligible.

(7a) Leibniz's rationalist objection to Newton's Theory of Gravitation exemplifies the radical version. He complained that if gravity were thought as a real force, then its effect would be a mysterious action at a distance. Leibniz blamed Newton for introducing "occult" forces into science, and until the end of his life Newton hoped to produce a properly "intelligible" account of gravity involving only action by contact interactions—he did not succeed. Modern scientific theories do not provide radical intelligibility. Once Galileo gave up his initial hope of presenting inertial motion as uniform circular motion, the theory of free fall he accepted left open at least as many whys as it closed. Why or how Galileo's mysterious mathematical structures arise in nature? The same goes for subsequent theorizing. Why or how the regularity given as Newton's law of gravitation arises? Why or how Fresnel's Core arise? Why or how the speed of light is a universal invariant? Contemporary fundamental theories fail radical intelligibility just as clearly.

Realists need not worry about this. Calls for radical intelligibility rest on views of cognition now widely recognized as mythical. Barring mystical insight and such, all actual understanding comes with opaque spots. At every scientific stage scientific warrant (and intelligibility) stops somewhere, albeit usually not at the traditional empiricist boundaries. Realism is compatible with suspending judgment about whether a certain theoretical claim correctly describes a fundamental or derivative aspect of nature. This is exemplified in the stance realists take towards e.g. Fresnel's Core, the invariance of light's speed, and fundamental principles in general.

A theory that saves all the known phenomena but whose reliable parts comprise only structures and explanations at phenomenal levels, provides the lowest level of understanding. This makes for a constructive empiricist take, which escapes skepticism by accepting realism about just the theory's empirical substructures. The point here is that radical theoretical intelligibility is not necessary for taking a realist stance towards a theory. From a selectivist

perspective, the key factor for taking a theory-part realistically is not the "intelligibility" it confers but its indispensability for maintaining the theory's predictive power in the context of current *background knowledge.* Ptolemaic orbits were denied realist interpretation not primarily because they failed the intelligibility requirement—Ptolemaic constructions went out of their way to honor, of all requirements, *intelligibility* (then guided by the Principle of Uniform Circular Motion for heavenly bodies and the Aristotelian arguments for the fixity of the Earth). Rather, Ptolemaic orbits were refused realist interpretation because the epicycles, deferents and equants they invoked were grossly *underdetermined by extant knowledge* (i.e. available data and cosmological principles). Positive evidence for the orbits specifically proposed was lacking.

None of this is not to question the realist relevance of theories that seek to achieve deep understanding. What is denied is that *scientific* realism must embrace radical intelligibility. Radical intelligibility is a trait realism about observables and every day affairs neither honors nor is expected to honor.

(7b) This brings us to cogent versions of the moderate intelligibility condition. Selectivists take a realist stance only towards theory-parts deemed to be both indispensable for the theory's success and free of compelling specific doubts against them (@@@). That is, the realist stance goes *only* to tenets for which there is strong positive evidence by modern scientific standards. In all the cases highlighted by realists, the selections supported by the strongest level of evidence available make the target domain intelligible well beyond the observable levels. When, by contrast, the positive evidence for a theory does not reach the unobservable explanatory posits that make the relevant phenomena intelligible, then the best stance to take about the theory is not realism but *constructive empiricism.* This clarifies what introductory characterizations of scientific realism get right about the intelligibility condition: A good theory must not have just significant predictive power but must also make the relevant phenomena *intelligible* (Richard DeWitt 2010: 72). If the theory parts that do this lack evidential warrant, then the reasonable stance towards them is constructive empiricism.

(8) **Realism Worth having.** Topping the above assumptions, there is a popular notion to the effect that a realist stance failing to adhere to most of the above requirements is "*not a realism worth having*". Against this idea, I have argued that none of the listed assumptions is worth

having. Every one of them lacks convincing warrant. Moreover, even if the assumptions did get proper warrant they face a deeper problem: the assumptions are *irrelevant* to the current realism/antirealism debate—they do not expose relevant contrasts between inferences limited to the phenomenal level and inferences that reach into theoretical levels.

In modern science, virtually all interesting augmentative inferences violate the listed assumptions. So, the latter simply and arbitrarily raise the epistemological standards of acceptability against theoretical assertions. If the above considerations are correct, then, realists and antirealists should reject the assumptions examined in this paper—they all rest on counterproductive myths and confusions.

References

Bohm, David (1957). *Causality and Chance in Modern Physics*. London: Routledge & Kegan Paul Ltd.

Devitt, Michael (2005). "Scientific Realism". In *The Oxford Handbook of Contemporary Philosophy*, Frank Jackson and Michael Smith, eds. Oxford: Oxford University Press: 767-91.

Giere, Ronald N. (2006). *Scientific Perspectivism*. Chicago: University of Chicago Press.

DeWitt, Richard (2010): *Worldviews*. Malden, MA: Wiley-Blackwell.

Kitcher, Philip, 1993. The Advancement of Science. Oxford: Oxford University Press.

Laudan, Larry. 1981. "A Confutation of Convergent Realism". *Philosophy of Science* 48: 19-49.

___. 1984. *Science and Values*. Berkeley: University of California Press.

___. 1996, *Beyond Positivism and Relativism: Theory, Method and Evidence*. Boulder, CO: Westview Press.

Leplin, Jarrett, ed. 1984. *Scientific Realism*. Berkeley: University of California Press.

___. 1997. *A Novel Defense of Scientific Realism*. New York: Oxford University Press.

Psillos, Stathis. 1999. *Scientific Realism: How Science Tracks Truth*. New York: Routledge.

Putnam, Hilary (1978).

Russell, Bertrand, 1931. *The Scientific Outlook*. London: George Allen & Unwin, Ltd.

Saatsi, Juha (2005). "Reconsidering the Fresnel-Maxwell Case Study." *Studies in History and Philosophy of Science* 36 (3): 509–38.

Saatsi, Juha and Peter Vickers (2011). "Miraculous Success? Inconsistency and Untruth in Kirchhoff's Diffraction Theory." *British Journal for the Philosophy of Science* 62: 29–46.

Stanford, P. Kyle. 2006. *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford: Oxford University Press.

Teller, Paul (2015). "Language and the Complexity of the World;" forthcoming.

Van Fraassen (1980). The Scientific Image. Oxford: Oxford University Press.

Vickers, Peter (2013). "A Confrontation of Convergent Realism". *Philosophy of Science* 80: 189-211.

Votsis, Ioannis (2011). "Saving the Intuitions: Polylithic Reference." Synthese 180 (2): 121–37.

Worrall, J. 1989a. "Fix it and be Damned: A Reply to Laudan". *British Journal for the Philosophy of Science* (40): 376-388.

------ (1989b). "Structural Realism: The Best of Both Worlds". *Dialectica* 43: 99-124.

------ (2016). "Structural Realism – the Only Viable Realist Game in Town." Forthcoming in *Scientific Realism: Objectivity and Truth in Science,* Wenceslao Gonzalez & Evandro Agazzi, (eds.).

Concrete Models and Holistic Modelling*

Wei Fang$^{\phi}$

Department of Philosophy, University of Sydney

Abstract: This paper proposes a holistic approach to the model-world relationship, suggesting that the model-world relationship be viewed as an *overall structural fit* where one organized whole (the model) fits another organized whole (the target). This approach is largely motivated by the implausibility of Michael Weisberg's weighted feature-matching account of the model-world relationship, where a set-theoretic conception of the structures of models is assumed. To show the failure of Weisberg's account and the plausibility of my approach, a concrete model, i.e. the San Francisco Bay model, is discussed.

* Draft paper, please do not quote without permission.

$^{\phi}$ Address: University of Sydney, NSW 2006, Australia. Email: wfan6702@uni.sydney.edu.au.

1. Introduction

One philosophical interest in the philosophy of modelling focuses on the problem of the model-world relationship, also known as the representation problem. Among many approaches to this problem, the similarity account has attracted much attention recently. Ronald Giere (1988, 1999a, 1999b, 2004, 2010), Peter Godfrey-Smith (2006) and Michael Weisberg (2012, 2013) have made the most substantial contributions.

The core of this account, first developed by Giere, is a view of the model-world relationship:

> The appropriate relationship, I suggest, is *similarity*. Hypotheses, then, claim a *similarity* between models and real systems. But since anything is similar to anything else in some respects and to some degree, claims of similarity are vacuous without at least an implicit specification of relevant *respects and degrees*. The general form of a theoretical hypothesis is thus: Such-and-such identifiable real system is similar to a designated model in indicated respects and degrees. (Giere 1988, 81; author's emphasis)

However, critics point out that this account is only schematic since it falls short of specifying the relevant *respects* and *degrees* (Suárez 2003). Moreover, Giere argues that a philosophical account of scientific representation should also take into consideration factors such as the *roles* played by scientists, and the *intentions* those scientists have when modelling (Giere 2004, 2010). Given these considerations, Weisberg develops a more sophisticated similarity account, called the *weighted feature-matching* account

(2012, 2013). The basic idea of his account comes from psychologist Amos Tversky's *contrast* account of similarity, which states that the similarity of objects *a* and *b* depends on the features they share and the features they do not. In light of this, Weisberg proposes his own account:

$$S(m, t) = \frac{\theta f(M_a \cap T_a) + \rho f(M_m \cap T_m)}{\theta f(M_a \cap T_a) + \rho f(M_m \cap T_m) + \alpha f(M_a - T_a) + \beta f(M_m - T_m) + \gamma f(T_a - M_a) + \delta f(T_m - M_m)} \quad (1)$$

f(x) refers to the weighting function, α, β, γ, δ, θ, and ρ denote weighting terms (parameters), subscripts *a* and *m* stand for attributes and mechanisms,[1] and *M* denotes the model and *T* the target. $(M_a \cap T_a)$ stands for attributes shared by the model and the target, $(M_a - T_a)$ attributes that the model has while the target does not, and $(T_a - M_a)$ attributes that the target has while the model does not. The same story goes for mechanisms *m*. Attributes and mechanisms as a whole are called *features* of the model and the target.

An interpretation for this equation is needed. First, there must be a feature set Δ, and the set of features of the model and the set of features of the target are defined as sets of features in Δ. The elements of Δ are determined by a combination of context, conceptualization of the target, and the theoretical goals of the scientist. Besides, the

[1] Properties and patterns of systems are termed attributes, and the underlying mechanisms generating these properties and patterns are termed mechanisms (Weisberg 2013, 145).

contents of Δ may change through time as science develops, which in turn might result in a reevaluation of the established model-world relationship (*Ibid.*, 149).

Second, consider the values of weighting parameters α, β, γ, δ, θ, and ρ. On Weisberg's account, different kinds of modelling require different weighting parameters. For example, if what interests us is the *minimal modelling* which concerns merely the mechanism responsible for bringing about the phenomenon of interest, the goal of this modelling is written as:[2]

$$\frac{|M_m \cap T_m|}{|M_m \cap T_m| + |M_a - T_a| + |M_m - T_m|} \rightarrow 1 \qquad (2)$$

Finally, consider the weighting function $f(x)$, telling us the relative importance of each feature in the set Δ. Weisberg says scientists in most cases have in their mind some subset of the features in Δ, which they regard as especially important. Hence some features are weighted more heavily, while others are equally weighted. Besides, the background theory determines which features in Δ should be weighted more heavily. If the background theory is not rich enough, deciding which should be weighted more heavily is partly an empirical problem.

Having presented an outline of Weisberg's account, I will now argue that this account fails to capture the relationship between concrete models and their targets. To illustrate this

[2] Weisberg also describes three other kinds of modelling requiring different weighting parameters: hyperaccurate, how-possibly and mechanistic modelling (2013, 150-52).

shortcoming (Sec. 3), I will first describe the San Francisco Bay model (Sec. 2). Sec. 4 will propose a holistic alternative to Weisberg's account, suggesting that the model-world relationship be viewed as an *overall structural fit* where one organized whole fits another organized whole. Sec. 5 will examine a case where the organization of the whole can be treated as simply another feature.

2. The San Francisco Bay Model

John Reber worried about the fragility of the water supply in the San Francisco Bay area in the 1950s. To solve this problem, he proposed an ambitious proposal, namely, to dam up the Bay. Carrying out this plan would not only supply San Francisco with unlimited drinking water but also entirely change the area's transportation, industrial, military and recreation landscape (Weisberg 2013, 1). However, his critics worried that Reber's plan would only achieve its aims at the cost of destroying commercial fisheries, rendering the South Bay a brackish cesspool, creating problems for the ports of Oakland, Stockton, and Sacramento, and so on (Jackson and Peterson 1977; Cf. Weisberg 2013, 1).

To settle this dispute, the Army Corps of Engineers was charged with investigating the overall influence of the Reber plan by building a massive hydraulic scale model of the Bay (Weisberg 2013, 1-2). Once the model was built, it was adjusted to accurately reproduce several measurements of the parameters such as tide, salinity, and velocities actually recorded in the Bay (for details see Army Corps of Engineers 1963). After the adjustment, it was time to verify the model:

> Agreement between model and prototype for the verification survey of 21-22 September 1956, and for other field surveys, was excellent. Tidal elevations, ranges and phases observed in the prototype were accurately reproduced in the model. Good reproduction of current velocities in the vertical, as well as in the cross section, was obtained at each of the 11 control stations in deep water and at 85 supplementary stations. The salinity verification tests for the verification survey demonstrated that for a fresh-water inflow into the Bay system […], fluctuation of salinity with tidal action at the control points in the model was in agreement with the prototype (Huggins and Schultz 1967, 11).

After the verification, modellers were in a good position to assess the Reber plan through the model built. The investigation showed that it would considerably reduce water-surface areas, reduce the velocities of currents in most of South San Francisco Bay, reduce the tidal discharge through the Golden Gate during the tidal cycle, and so forth (Huggins and Schultz 1973, 19). Given these disastrous consequences, the Army Corps then denounced Reber's plan (Weisberg 2013, 9).

3. How Could Weisberg's Account Shed Light on the Bay Model?

I have argued elsewhere that Weisberg's account cannot shed light on mathematical models due to its atomistic conception of features and its assumption of the set-theoretic approach to model structures (citation anonymized). I find that the same charges can be raised in the case of concrete models.

Consider the first charge: Weisberg's account is committed to an atomic conception of features. The key of Weisberg's account is the claim that the similarity of objects *a* and *b* depends on the features they share and the features they do not share. Let us take a closer look at the equation (1). The numerator invites us to weight features shared, and the denominator asks us to weight all features involved (including three feature subsets: features shared, features possessed by the model but not the target, and features possessed by the target but not the model). Each feature is weighted independently and only once, with it falling into one of the three feature subsets. The numerator is the weighted sum of features shared, the denominator is the weighted sum of features shared and unshared, and the similarity measure is the ratio of the numerator to the denominator.

However, features in the Bay model are not atomistic and independent of each other. As Huggins and Schultz put it explicitly, "Among the problems to be considered were the conservation of water […]; […] the tides, currents and salinity of the Bay as they affect other problems […]. None of these problems can be studied separately, for each affects the others" (1973, 12). The reason why none of these problems can be studied separately is because factors involved in these problems cannot be studied separately.

Consider, for instance, the relationship between two key features in the model: tide and salinity. Salinity levels vary along an estuary depending on the mixing of freshwater and saltwater at a site. An estuary "is the transition between a river and a sea. There are two main drivers: the river that discharges fresh water into the estuary and the sea that fills the estuary with salty water, on the rhythm of the tide" (Savenije 2005, Preface ix).

To illustrate this "rhythm of the tide", consider the effect of the spring-neap tidal cycle on the vertical salinity structure of the James, York and Rappahannock Rivers, Virginia, U.S.A.:

> Analysis of salinity data from the lower York and Rappahannock Rivers (Virginia, U.S.A.) for 1974 revealed that both of these estuaries oscillated between conditions of considerable vertical salinity stratification and homogeneity on a cycle that was closely correlated with the spring-neap tidal cycle, i.e. homogeneity was most highly developed about 4 days after sufficiently high spring tides while stratification was most highly developed during the intervening period. (Haas 1977, 485)

This short report shows not only that characteristics of salinity (such as stratification and homogeneity) are influenced by characteristics of the tide, but also that there is a phase connection (or synchronization) between tidal cycle and salinity oscillations. The former is a causal relationship while the latter is a temporal relationship. The phase connection among features was also emphasized by the Army Corps when verifying the Bay model, saying "These gages were installed in the prototype and placed in operation several months in advance of the date selected to collect the primary tidal current and salinity data required for model verification, since *it was essential to obtain all data simultaneously for a given tide over at least one complete tidal cycle of 24.8 hours*" (1963, 50; my emphasis). Moreover, the same story goes for tide and tidal currents (for details see Army Corps 1963, 20).

In short, features in a model bear not only causal relationships, but also temporal relationships to one another. This implies that, when verifying the model, features of the

model causally interact with each other in producing certain outputs (e.g. predictions, effects, phenomena, etc.), rather than that they individually or separately produce outputs. So although outputs of key features in the Bay model can be identified and measured separately, they are not produced separately.

It is important to note that the causal interaction among features may lead to a different kind of interaction, i.e. a "similarity interaction",[3] wherein features interact with one another in producing the similarity value. That is, one feature's contribution to the similarity value depends on other feature(s)' contribution to that value.[4] The difference between causal and similarity interaction is that the latter is a statistical relationship among measured features, and can be viewed as a reflection of the former when coupled with an assumption that there might be such an underlying causal structure.[5] For example, a similarity interaction is shown by the verification of salinity in the Bay model, where the measurement of salinity (as a measurement of one feature's contribution to the similarity

[3] I thank X for suggesting this term for me.

[4] This point can be best illustrated with the curve fitting example: when computing the fit of a straight line $y=ax+b$ to a cloud of points, a and b will depend on each other to produce the best fit (I thank X for giving me this example).

[5] This assumption is important because there are cases where the fact that there is similarity interaction cannot guarantee that there is also causal interaction, because some randomly generated data set may also show interaction among features. In other words, causal interaction can lead to similarity interaction and the reverse is not true (I thank Y for letting me know this). I will discuss this assumption, called "precondition" later, in Sec. 4.

value from Weisberg's perspective) depended on other features in the way in which other features were kept constant: "salinity phenomena in the model were in agreement with those of the prototype *for similar conditions of tide, ocean salinity, and fresh-water inflow*" (*Ibid*., 54; my emphasis).

The way that similarity interaction reflects causal interaction, when coupled with the assumption mentioned above, can be expressed as follows: if what is under verification is a causal structure to which modellers do not have direct access (so the structure cannot be a feature in Weisberg's formula), then the coherent behavior of features (i.e. their similarity interactions such as phase connections) is a way of verifying, or at least indicating, the causal interactions in the underlying causal structure.[6] That is the reason why it was so essential to obtain all data simultaneously within a complete tidal cycle for the Bay model, and why all other features must be kept constant when verifying salinity (or other features).

Given features' causal interactions in the model and their similarity interactions when measuring them, it seems that assessing the relationship between a model and its target cannot be simply achieved in the way suggested by Weisberg's equation, for features' contribution to the similarity relationship is not *additive* but *interactive*. That is, to assess the relationship between a model and its target, one cannot measure each feature's contribution independently and then add them together.

4. Set-Theoretic or Non-Set-Theoretic? A Holistic Alternative

[6] I thank X for bringing this point to my attention.

Now we arrive at the problem of why Weisberg's account is deeply committed to an atomistic conception of features. As I have argued elsewhere, this problem ultimately comes down to Weisberg's understanding of the structure of models (citation anonymized). Weisberg says models are *interpreted structures* (2013, 15), so concrete models are interpreted concrete structures. At first glance, I have no quarrel with this understanding. On closer inspection, however, it can be shown that Weisberg's account on the model-world relationship assumes a set-theoretic approach to the structure of models.[7] This is because Weisberg's similarity measure can be derived from the *Jaccard similarity coefficient* between two sets, a coefficient assuming a set-theoretic conception of objects (citation anonymized).

The key to the set-theoretic approach to structures is its assumption that elements of objects (i.e. models and targets) are independent of each other, just as elements of a set are independent of each other. In other words, it construes both the model and the target as a set of independent elements, the similarity between which consists in the ratio of the number of elements shared to the number of all elements (citation anonymized). However, as discussed in Sec. 3, features are not independent. More importantly, their causal interactions may result in a similarity interaction among features.

This similarity interaction undermines Weisberg's account, for it cannot properly capture the dependence relationship of features' contribution to the overall similarity

[7] Note that Weisberg *explicitly* objects to the set-theoretic approach to models (2013, 137-42). However, I think it is compatible to claim that someone *implicitly* assumes what someone explicitly rejects.

measure between a model and a target. Nonetheless, there is still a way to save the very intuitive notion of similarity, by abandoning the set-theoretic conception of structures. That is, if the structure of a model is viewed as an *organized whole* in which each component of the whole is interconnected to other component(s) (directly or indirectly) in such a way that they interact with one another in producing certain phenomena of interest (i.e. outputs). Under such an understanding, therefore, assessing the relationship between a model and its target cannot be simply achieved by assessing each individual feature's relationship and then adding them together. Nor can this be done by assessing each connection among two or more features and then adding them together, even if connections (causal or non-causal) are also interpreted as features. On the other hand, however, the notion of similarity can be minimally preserved by claiming that assessing the similarity or *fit* (I will use fit hereafter) between a model and a target amounts to assessing the *overall structural fit* between the model and its target.

Generally speaking, structural fit means the structure of the model fits the structure of the target *as an organized whole*. That said, nevertheless, it should be stressed that there is no univocal meaning for the term "structural fit" that could encompass all circumstances, nor can a single equation or formula capture all situations. This is largely due to the heterogeneity of modelling practice and its multifarious goals. On the other hand, however, instructive points can still be asserted. In what follows I will elaborate some basics regarding the conception of "structural fit".

Structural fit in mathematical modelling means different things than in concrete modelling. For example, in a very simple case of curve fitting where a straight line $y=ax+b$

is fitted to a cloud of points, features *a* and *b* will interact with each other to produce the best fit. That is, what fits the cloud of points is the overall structure, not the additive sum of each individual feature. As I have argued elsewhere, in more complicated mathematical modelling such as the *maximum likelihood estimation*, the fit is usually achieved through comparing the predicted data set derived from the model *as a whole* to the observed data set derived from the target system (citation anonymized). Individual features of the model simply disappear, and causally related features, as constituting a whole, that co-occur in the data set are what really matters.

In the case of concrete modelling, admittedly, the claim that assessing the fit between a model and a target amounts to assessing the overall structural fit seems to be less apparent. On closer examination, however, the same claim still holds. Let us go back to the verification of the Bay model. At first glance, it seems the verification of the model was achieved by independently verifying the output (i.e. data sets) of each individual feature, as the report showed (see Sec. 2 for the verification report). That is, it seems that by verifying that each feature in the model fits its counterpart in the target, scientists made the judgment that the model fits the target system.

Underlying this seemingly plausible reasoning, however, there remains the problem of why we are allowed to confirm the verification of the model by means of only verifying several outputs of individual features. Or, to put it slightly differently, in terms of what does the fit of features guarantee the judgment about the fit of the model to the target? I take it that it is more than the fit of individual features themselves that makes sense of the reasoning that the model fits the target. There must be a precondition for this reasoning

(remember the "assumption" made in the last section). After all, there are many cases in which the fit of features does not guarantee the fit of the model itself to the target. For instance, a drawing of Tom's face may accurately capture all features of his face, e.g., nose, eyes, mouth, etc., but still falls short of fitting his face, because of the wrong organization of these features, e.g., putting the mouth in between the eyes and nose (Weisberg would argue that the organization could be a feature. I will discuss this point in Sec. 5.).

So if the fit of features is insufficient to vindicate the fit of a model to its target, what could provide this vindication? My claim is, contrary to Weisberg, that it is the *overall structural fit* of the model to the target system that warrants the fit judgment about the model and its target. In other words, the fit of individual features can only succeed in supporting the fit of the model to the target by the precondition that these features can be organized into the whole (i.e. the assumption that there is such an underlying causal structure), not the other way around.

To understand this "holistic reasoning", let me articulate the specifics involved step by step. We first build a concrete model, i.e. a concrete structure, wherein features are interconnected with one other in such a way that they have the potential to interactively produce certain phenomena of interest (i.e. outputs). Before verifying the model, we need to adjust key features to make sure the model works very well. Note that any adjustment will not simply be the adjustment of individual features but also of their interconnections, resulting in the adjustment of the overall structure of the model. Finally, we verify the model by comparing the outputs of the model to the outputs of the target. As with mathematical models, this verification is also usually made via comparing data sets, as

shown in the Bay model. Note that though these outputs can be identified, derived and measured independently, it is causally connected features that interact in producing them. In other words, although you verify each feature separately, the support provided by a single feature is not confined to that feature of the model, but confirms all aspects of the model that are involved in generating that output.

Thus understood, therefore, the gist of verifying a concrete model such as the Bay model can be captured as follows. The verification of each feature, as a component of a whole, is simply the verification of one aspect of the structure. So the verification of different features is the verification of the same structure from different perspectives. Thus, if the model is an organized whole, then the more features that are independently verified the more likely it is that the model resembles the reality. On the other hand, if what is under verification is not an organized whole but an aggregation of independent items, then the verification of each lends no credence to other parts of the aggregated whole—because these items are not causally linked, the verification of each item is only the verification of that item itself.

In sum, the relationship between a concrete model and its target is a holistic matter wherein an organized whole fits (to a certain degree) or fails to fit another organized whole. Though it seems at first blush that the verification of the whole results from the sum of the verification of each component, the real picture is just the reverse: the whole is always in place and the component can gather force in supporting the verification of the whole only when it can be organized into the whole.

5. Organization and Features

As mentioned above, Weisberg would argue that the organization could be a feature, so a drawing of Tom's face capturing accurately not only his nose, mouth, eyes but also their organization can be a good model of Tom's face. A holistic account agrees that organization could be a feature, but disagrees with the way that organization is treated in Weisberg's similarity measure. Intuitively, we may say that a drawing of one person's face is a good model if it has the right features: such as a nose, a mouth, eyes, and the organization of all of these. So it seems that if you get each individual feature right, then you get the whole model right. That is, features *additively* contribute to the goodness of the model.

This intuitive way of understanding scientific modelling, however, obscures the fact that features may interact in producing the fit of a model, as shown in Sec. 4. To reiterate this point and to draw a connection to our current discussion, consider another ordinary example.[8] Suppose Anne's face is an ideal one which scientists want to model. Anne has an ideal nose, which is straight, in contrast to a non-ideal nose, which might be bumped or concave. She also has an ideal nostril, which is round, in contrast to a non-ideal one, which might be triangular or square. Scientist A draws a face for Anne that has a round nostril and a concave nose, while scientist B draws a face that has a triangular nostril and a bumped nose. Drawing A has an ideal feature (the round nostril), but neither feature of drawing B is ideal. Now we ask which drawing better fits Anne's face. It is likely that we

[8] I thank X for giving me this nice example.

will say that B is better because our contemporaries' taste tells us that there is no face so ugly as one with a round nostril and a concave nose, though a round nostril itself is ideal. Hence we see a case wherein the nostril and nose interact to produce the fit of a model to a target.

This discussion leads to a more general question: what are features? In Weisberg's account, a model can *more or less* fit a target, but features are either shared or not. Yet as Wendy Parker points out, "relevant similarities often seem to occur at the level of individual features, not just at the level of the model" (2015, 273). This is because features themselves can be objects such that they more or less fit each other.[9] Weisberg may argue that this problem can be fixed by the assumption that a feature can be redescribed as a set of sub-features, so the similarity between two features can be measured as the result of the similarity between their sub-features. However, I see this treatment as a non-starter, for the similarity between sub-features may also be a matter of degree such that it should be measured as the result of the similarity between their sub-sub-features, and between their sub-sub-sub-features, and so on.

On the other hand, a holistic account does not encounter this problem: if a feature is an object, then it can be viewed as an organized whole. So the relationship between a feature in a model and a feature in a target also consists in their structural fit. Take a minimal model for instance. Most minimal models primarily attempt to represent repeatable patterns of behavior largely insensitive to underlying microscopic details (Batterman 2002, 27). Suppose we are interested in the buckling behavior of struts, and write a

[9] I thank X for bringing this to my attention.

phenomenological formula, called Euler's formula, to characterize it (see Batterman 2002 for details). It seems the pattern of behavior is the only feature involved in this case, i.e., a dependence relationship among several parameters. So assessing the fit between the model and the target comes down to assessing the fit between the feature in the model and the feature in the target. For this, a holistic account can easily come through: the relationship is an overall structural fit, wherein a dependence relationship as a feature fits another dependence relationship.

6. Conclusion

This paper has shown that the assumption of a set-theoretic approach to structures makes Weisberg's account fail to shed light on the San Francisco Bay model. Alternatively, a holistic approach to models, viewing the model-world relationship as an overall structural fit, fares better not only in capturing the Bay model, but more generally in making sense of modelling practice.

References

Army Corps of Engineers. 1963. *Technical Report on Barriers: A Part of the Comprehensive Survey of San Francisco Bay and Tributaries, California*. Appendix H, Volume 1: Hydraulic Model Studies. San Francisco: Army Corps of Engineers.

Batterman, Robert. 2002. "Asymptotics and the Role of Minimal Models." *British Journal for the Philosophy of Science* 53 (1): 21-38.

Giere, Ronald N. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Giere, Ronald N. 1999a. *Science without Laws*. Chicago: University of Chicago Press.

Giere, Ronald N. 1999b. "Using Models to Represent Reality." In *Model-Based Reasoning in Scientific Discovery*, ed. Lorenzo Magnani, Nancy J. Nersessian, and Paul Thagard, 41-57. Springer Science & Business Media.

Giere, Ronald N. 2004. "How Models Are Used to Represent Reality." *Philosophy of Science* 71 (5): 742-752.

Giere, Ronald N. 2010. "An Agent-Based Conception of Models and Scientific Representation." *Synthese* 172 (2): 269–281.

Godfrey-Smith, Peter. 2006. "The Strategy of Model-based Science." *Biology and Philosophy* 21 (5): 725-740.

Haas, Leonard W. 1977. "The Effect of the Spring-Neap Tidal Cycle on the Vertical Salinity Structure of the James, York and Rappahannock Rivers, Virginia, U.S.A." *Estuarine and Coastal Marine Science* 5:485-496.

Huggins, Eugene. M., and Edward A. Schultz. 1967. "San Francisco Bay in A Warehouse." *Journal of the IEST* 10 (5): 9-16.

Huggins, Eugene M., and Edward A. Schultz. 1973. "The San Francisco Bay and the Delta Model." *California Engineer* 51 (3): 11-23.

Jackson, W. Turrentine, and Alan M. Peterson. 1977. *The Sacramento-San Joaquin Delta: The Evolution and Implementation of Water Policy*. Davis: California Water Resource Center, University of California.

Parker, Wendy. 2015. "Getting (even more) serious about similarity." *Biology and Philosophy* 30 (2): 267-276.

Savenije, Hubert H. G. 2005. "*Salinity and Tides in Alluvial Estuaries*." Elsevier Science.

Suárez, Mauricio. 2003. "Scientific Representation: against Similarity and Isomorphism." *International Studies in the Philosophy of Science* 17 (3): 225-244.

Weisberg, Michael. 2012. "Getting Serious about Similarity." *Philosophy of Science* 79 (5): 785-794.

Weisberg, Michael. 2013. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

PROBABILISTIC ACTUAL CAUSATION

LUKE FENTON-GLYNN

DEPARTMENT OF PHILOSOPHY, UNIVERSITY COLLEGE LONDON

GOWER STREET, LONDON, WC1E 6BT, U.K.

ABSTRACT. Actual (token) causes – e.g. Suzy's being exposed to asbestos – often bring about their effects – e.g. Suzy's suffering mesothelioma – probabilistically. I use probabilistic causal models to tackle one of the thornier difficulties for traditional accounts of probabilistic actual causation: namely probabilistic preemption.

Luke Fenton-Glynn

1. INTRODUCTION

Actual (token) causation is the relation that obtains when, for example, Suzy's being exposed to asbestos causes her to suffer mesothelioma. A number of theorists (e.g. Halpern and Pearl 2001, 2005; Hitchcock 2001, 2007; Weslake 2016) have deployed structural equations models (SEMs) in developing novel solutions to difficulties confronting traditional accounts of this relation. These theorists have focused on *deterministic* actual causation (DAC).[1] I draw on probabilistic causal models (PCMs) – analogues of deterministic SEMs – to provide an account of probabilistic actual causation (PAC). I don't attempt to show that my account can handle the full battery of test cases discussed in the literature. I simply demonstrate that it yields an elegant treatment of one very central case – probabilistic preemption – with a view to motivating further investigation of formal approaches to PAC.

2. PROBABILITY-RAISING

Probability-raising is central to the account developed here – as on traditional accounts of PAC.[2] To explain how I will understand that notion a bit of stage-setting is required.

I take the relata of the actual causal relation to be variable values. Adopting Goldszmidt and Pearl's (1992, 669–70) notation, $P(W = w|do(V = v))$ represents the probability for $W = w$ that *would* obtain if V were set to $V = v$ by an 'intervention' (Woodward 2005, 98). This is liable to diverge from the conditional probability $P(W = w|V = v)$: witness the difference between the probability of a storm *conditional* upon the barometer needle pointing toward the

[1] Cf. Halpern and Pearl (2005, 852); Hitchcock (2007, 498).

[2] Reichenbach (1971, 204); Suppes (1970); Lewis (1986, 175–84); Menzies (1989). The deficiencies of these accounts have been demonstrated by e.g. Salmon (1984, 192–202); Menzies (1996, 85–96); Hitchcock (2004).

word 'storm' and the probability of a storm if I had intervened upon the barometer needle to point it toward 'storm'.

Variable X taking value $X = x$ (rather than $X = x'$) raises the probability of $Y = y$ in the relevant sense iff:[3]

$$P(Y = y|do(X = x)) > P(Y = y|do(X = x')) \tag{1}$$

Appealing to interventionist probabilities means avoiding probability-raising relations between independent effects of a common cause, such as the barometer reading and the storm (cf. Lewis 1986, 178).

Probabilistic preemption cases illustrate that straightforward probability-raising is neither necessary nor sufficient for causation (Menzies 1989, 1996).

3. Probabilistic Preemption

The following example is inspired by Anscombe (1971).[4]

[3]Here and throughout, the probabilities (chances) should be taken to be those obtaining immediately after the interventions bringing about the variable values specified in the scope of the $do(\cdot)$ function have occurred (cf. Lewis 1986, 177).

[4]The probabilities involved (except the decision probabilities) are quantum and therefore objective and able underwrite causal relations. (If you're worried that the decision probabilities are not objective, the example could be complicated so that the decisions are made on the basis of outcomes of quantum measurements.) I find it plausible that the probabilities of many high level sciences are also objective (cf. e.g. Loewer 2001; Ismael 2009).

Luke Fenton-Glynn

(ProbPre) *Someone (neither you nor I) has connected a Geiger counter to a bomb so that the bomb will explode if the Geiger registers above a threshold reading. I place a place a chunk of U-232 (half-life = 68.9 years; decays by α-emission) near the Geiger. By chance, enough U-232 atoms decay within a short enough interval for the Geiger to reach the threshold reading so that the bomb explodes. Unbeknownst to me, you've been standing nearby observing. You have a chunk of Th-228 (half-life = 1.9 years; decays by α-emission), which contains many more atoms than my chunk of U-232. You've decided that you'll place your Th-228 near the Geiger iff I fail to place my U-232 near the Geiger. There's a negligible chance that you won't follow the course of action you've decided on. Seeing that I place my U-232 near the Geiger, you don't place your Th-228 near the Geiger.*[5]

Let M, D, Y, T, and E be binary variables which, respectively, take value 1 if the following things occur (and 0 otherwise): I place my U-232 near the Geiger; you decide to place your Th-228 near the Geiger iff I don't place my U-232 near the Geiger; you place your Th-228 near the Geiger; the threshold reading is reached; the bomb explodes.

My act ($M = 1$) was an actual cause of the explosion ($E = 1$). Yet plausibly the following inequality holds:

$$P(E = 1|do(M = 1)) < P(E = 1|do(M = 0)) \tag{2}$$

[5]The range of α-particles is 3-5 cm. Suppose that, for each of us, a decision to place our chunk 'near' the Geiger counter is a decision to place it $<$ 5cm away and a decision not to place it nearby is a decision to place it nowhere near ($\gg$ 5cm away).

That is, my placing my U-232 near the Geiger *lowers* the probability of the bomb exploding because it strongly lowers the probability of your placing your more potent Th-228 near the Geiger. Probability-raising is therefore unnecessary for actual causation.

Your decision ($D = 1$) was *not* an actual cause of the explosion, since you don't place your Th-228 near the Geiger. Yet provided there's some chance that $M = 0$, the following inequality holds:

$$P(E = 1|do(D = 1)) > P(E = 1|do(D = 0)) \tag{3}$$

Inequality (3) holds because your decision raises the probability that the bomb will still explode in the scenario in which $M = 0$.[6] Probability-raising is therefore insufficient for actual causation.

Actual causation therefore can't be identified with probability-raising. In developing a more nuanced analysis, it is helpful to appeal to PCMs.

4. PCMs

A PCM, $\mathscr{M}$, is a 5-tuple $\langle \mathscr{V}, \mathscr{C}, \Omega, \mathscr{F}, do(\cdot) \rangle$. $\mathscr{V}$ is a set of variables. Suppose $\mathscr{R}$ denotes a function from elements of $\mathscr{V}$ to sets of values: for all $V \in \mathscr{V}$, $\mathscr{R}(V)$ is the *range* of V. In Halpern and Pearl's (2005, 851–2) terminology, a formula $V_i = v_i$, for $V_i \in \mathscr{V}$ and $v_i \in \mathscr{R}(V)$, is a *primitive event*. $\mathscr{C}$ is the set of all those possible conjunctions of primitive

[6] $D = 0$ is multiply realizable: there is more than one alternative to the decision that you in fact make. E.g. you could decide that you will place your Th-228 near the Geiger no matter what, or that you will not do so no matter what. We can stipulate that the latter alternative is much more probable.

events, $V_1 = v_1 \& \ldots \& V_n = v_n$, such that $V_i \in \mathscr{V}$ and $v_i \in \mathscr{R}(V_i)$ and such that, for no pair of conjuncts $V_i = v_i$, $V_j = v_j$ is $V_i \equiv V_j$, and where no two elements of $\mathscr{C}$ differ *only* in the permutation of their conjuncts. Such a conjunction is denoted $\mathbf{V} = \mathbf{v}$ (primitive events and the null event are limiting cases of such conjunctions). Abusing notation, the fact that $v_i \in \mathscr{R}(V_i)$ for each primitive event $V_i = v_i$ in the conjunction $\mathbf{V} = \mathbf{v}$, is abbreviated $\mathbf{v} \in \mathscr{R}(\mathbf{V})$ and the set of variables that appear in $\mathbf{V} = \mathbf{v}$ is denoted $\mathbf{V}$.

Call a conjunction $\mathbf{V} = \mathbf{v}$ *maximal* if it contains a conjunct of the form $V_i = v_i$ for each $V_i \in \mathscr{V}$. Ω is the set of all maximal conjunctions of primitive events. $\mathscr{F}$ is a sigma algebra on Ω. Finally, $do(\cdot)$ is a function from elements of $\mathscr{C}$ to probability distributions on $\mathscr{F}$ (cf. Pearl 2009, 70, 110): for each element $\mathbf{V} = \mathbf{v}$ of $\mathscr{C}$, $P(\cdot|do(\mathbf{V} = \mathbf{v}))$ is the probability (chance) distribution on $\mathscr{F}$ that *would* obtain if interventions were performed to bring about $\mathbf{V} = \mathbf{v}$.

A PCM can be represented graphically by taking the variables in $\mathscr{V}$ as nodes and drawing a directed edge from V_i to V_j ($V_i, V_j \in \mathscr{V}$) iff, where $\mathbf{S} = \mathscr{V} \backslash V_i, V_j$, there is some assignment of values $\mathbf{s}' \in \mathscr{R}(\mathbf{S})$, some pair of values $v_i, v_i' \in \mathscr{R}(V_i)$ ($v_i \neq v_i'$) and some value $v_j \in \mathscr{R}(V_j)$ such that $P(V_j = v_j|do(V_i = v_i \& \mathbf{S} = \mathbf{s}')) \neq P(V_j = v_j|do(V_i = v_i' \& \mathbf{S} = \mathbf{s}'))$.

In constructing a PCM, $\mathscr{M}_{Pre}$, of **(ProbPre)** we might take the variable set to be $\mathscr{V}_{Pre} = \{D, M, Y, T, E\}$. The range of each variable in $\mathscr{V}_{Pre}$ is the pair $\{0, 1\}$. $\mathscr{C}_{Pre}$, Ω_{Pre}, and $\mathscr{F}_{Pre}$ are generated by $\mathscr{V}_{Pre}$ and $\mathscr{R}_{Pre}$ in the way described above. For each element of $\mathscr{C}_{Pre}$, the function $do(\cdot)$ returns the chance distribution on $\mathscr{F}_{Pre}$ that would obtain if interventions were performed to bring about that element of $\mathscr{C}_{Pre}$. The graph for $\mathscr{M}_{Pre}$ is given as figure 1.

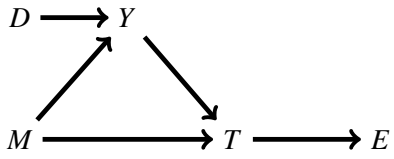


FIGURE 1

A directed path in a graph is an ordered sequence of nodes, $\langle V_1, V_2, \ldots, V_n \rangle$, such that there is a directed edge from V_1 to V_2, and a directed edge from V_2 to $\ldots V_n$. $\langle M, Y, T, E \rangle$ is an example of a directed path in the graph of $\mathscr{M}_{Pre}$.

5. Appropriate Models

In Section 6, I provide a definition of what it is for $X = x$ (rather than $X = x'$) to count as an actual cause of $Y = y$ *relative to a PCM*. I then define a non-model-relativized notion of actual causation by saying that $X = x$ (rather than $X = x'$) counts as an actual cause of $Y = y$ *simpliciter* provided that $X = x$ (rather than $X = x'$) counts as an actual cause $Y = y$ relative to at least one *appropriate* PCM.[7] A similar strategy is commonly adopted by those analyzing DAC in terms of SEMs (Hitchcock 2001, 287, 2007, 503; Weslake 2016). This requires an account of 'appropriate' models.

Many of the criteria for an appropriate SEM for evaluating DAC carry over to PCMs, including the following three:

(Partition) For all $V \in \mathscr{V}$, the elements of $\mathscr{R}(V)$ should form a partition (Halpern and Hitchcock 2010, 397–8; Blanchard and Schaffer 2016)

(Independence) For no two variables $V, W \in \mathscr{V}$ should there be elements $v \in \mathscr{R}(V)$ and $w \in \mathscr{R}(W)$ such that the states of affairs represented by $V = v$ and $W = w$ are logically or metaphysically related (Hitchcock 2001, 287; Halpern and Hitchcock 2010, 397)

[7] As the parentheses indicate I define a *contrastive* relation of actual causation. Where variables are binary – as in $\mathscr{M}_{Pre}$ – this is inconsequential and I will typically suppress such parentheses. But it becomes important in cases of multi-valued variables (see Halpern and Pearl 2005, 859).

(Naturalness) For all $V \in \mathscr{V}$, $\mathscr{R}(V)$ should include only values that represent reasonably natural and intrinsic states of affairs. (Blanchard and Schaffer 2016)

The analysis of actual causation proposed below takes all and only values of distinct variables to be potential causal relata. (Partition) insures that we don't thereby miss actual causal relations because they obtain between the values of a single variable. (Independence) insures that we don't mistake stronger-than-causal relations for causal relations. (Naturalness) insures that unnatural or non-intrinsic states of affairs do not get counted as causes and effects (see Lewis 1986, 190, 263; Paul 2000, 245).[8]

A further condition is that a model is appropriate for evaluating whether $X = x$ is an actual cause of $Y = y$ in world θ only if it satisfies (Veridicality):

(Veridicality) For any conjunction $\mathbf{V} = \mathbf{v} \in \mathscr{C}$ taken as an input, the probability distribution $P(\cdot|do(\mathbf{V} = \mathbf{v}))$ yielded as an output by $do(\cdot)$ should be the *objective chance* distribution over $\mathscr{F}$ that would$_\theta$ result from interventions setting $\mathbf{V} = \mathbf{v}$. ('Would$_\theta$' indicates that what is required is that this counterfactual be true in θ.)

(Veridicality) is an analogue – for PCMs – of the requirement that SEMs encode only true counterfactuals (Hitchcock 2001, 287, 2007, 503).

In the DAC/SEMs literature another condition on model appropriateness is typically added:

(Serious Possibilities) $\mathscr{V}$ should not be such as to generate elements of Ω that represent possibilities "that we consider to be too remote" (Hitchcock 2001, 287;

[8] If *absences* are unnatural states of affairs (cf. Lewis 1986, 189–93), we might instead require that each variable have *at most one value* representing such a state of affairs.

cf. Woodward 2005, 86–91, Weslake 2016, Blanchard and Schaffer 2016).

We likely need this requirement too. A discussion of whether the vagueness and subjectivity thereby introduced is problematic would take us too far afield.[9] Still, it doesn't put the present account in any *worse* shape than its deterministic analogues. Moreover, traditional accounts of actual causation – which don't appeal to causal models – also stand in need of appeal to 'serious possibilities' (Woodward 2005, 86–8).

A final requirement – similar to one imposed in the DAC/SEM literature – for a model $\mathscr{M}$ to be an appropriate one for evaluating whether $X = x$ is an actual cause of $Y = y$ in world θ is:

(Stability) There is no model $\mathscr{M}^*$ (satisfying Partition, Independence, Naturalness, Veridicality, and Serious Possibilities) with a variable set $\mathscr{V}^*$ such that $\mathscr{V}^* \supset \mathscr{V}$ relative to which $X = x$ (rather than $X = x'$) is *not* an actual cause of $Y = y$. (Halpern and Hitchcock 2010, 394–5; Blanchard and Schaffer 2016; Halpern 2014; Hitchcock 2007, 503).

The idea is that an appropriate model is a sufficiently rich representation of causal reality that moving to a richer representation would not reveal an apparent actual causal relation to be spurious.[10]

The converse requirement – that a negative verdict about actual causation should not be overturned in a richer model – isn't needed. This is because actual causation (simpliciter) is defined in terms of actual causation relative to *at least one* appropriate model. A model relative verdict that $X = x$ is not an actual cause of $Y = y$ thus automatically fails to translate

[9] See Woodward (2005, 86–91).

[10] (Stability) renders the notion of an appropriate model relative to the causal claim being evaluated.

into a verdict that $X = x$ is not an actual cause (simpliciter) of $Y = y$ if there is a richer (and otherwise appropriate) model relative to which $X = x$ is an actual cause of $Y = y$.

We can now state a definition of actual causation in terms of appropriate PCMs that handles (**ProbPre**).

6. PAC

Actual causation *simpliciter* is defined in terms of actual causation relative to an appropriate PCM. Model-relative actual causation is then defined.[11]

AC(S)
Where $x, x' \in \mathscr{R}(X)$ *and* $y \in \mathscr{R}(Y)$, $X = x$ *(rather than* $X = x'$*) is an* actual cause (simpliciter) *of* $Y = y$ *in world* θ *iff* $X = x$ *(rather than* $X = x'$*) is an actual cause of* $Y = y$ *relative to* at least one *model* $\mathscr{M}$ *(with* $X, Y \in \mathscr{V}$*) that is appropriate for evaluating whether* $X = x$ *(rather than* $X = x'$*) is an* actual cause (simpliciter) *of* $Y = y$ *in* θ.

[11] Those familiar with Halpern and Pearl's (2001, 2005) analyses of DAC are invited to see an analogy with **AC(M-R)**. **AC(M-R)** was partly inspired by thinking about how a counterpart of Halpern and Pearl's analysis might be developed that is adequate to the probabilistic case. Ultimately, I'm optimistic that an adequate account of DAC will fall out of an adequate account of PAC as the special case where all probabilities are 1 or 0. This is why my definitions take the definiendum to be 'actual cause' rather than 'probabilistic actual cause'.

Probabilistic Actual Causation

AC(M-R)

Where $x, x' \in \mathscr{R}(X)$ and $y \in \mathscr{R}(Y)$, $X = x$ (rather than $X = x'$) is an *actual cause* of Y = y relative to a model $\mathscr{M}$ (with $X, Y \in \mathscr{V}$) in world θ iff there is a partition $(\mathbf{Z}, \mathbf{W})$ of $\mathscr{V} \backslash X, Y$ and some setting $\mathbf{W} = \mathbf{w}'$ of the variables in $\mathbf{W}$ such that the $do(\cdot)$ function associated with $\mathscr{M}$ entails that, for all subsets $\mathbf{Z}'$ of $\mathbf{Z}$ (where, for each such subset, $\mathbf{Z}' = \mathbf{z}^*$ are the values that the variables in $\mathbf{Z}'$ have in θ):

(**IN**) $P(Y = y | do(X = x \& \mathbf{W} = \mathbf{w}' \& \mathbf{Z}' = \mathbf{z}^*)) > P(Y = y | do(X = x' \& \mathbf{W} = \mathbf{w}'))$

AC(M-R) counts $M = 1$ as an actual cause of $E = 1$ relative to $\mathscr{M}_{Pre}$ (and the world described in (**ProbPre**)). Consider the partition of $\mathscr{V}_{Pre} \backslash M, E$ such that $\mathbf{W} = \{D, Y\}$ and $\mathbf{Z} = \{T\}$. And consider the assignment $\{D = 1, Y = 0\}$ of values to the variables in $\mathbf{W}$. **AC(M-R)** is satisfied because (**IN**) holds for both subsets of $\mathbf{Z}$ ($\emptyset$ and $\{T\}$), as shown by (4) and (5):

(4) $P(E = 1 | do(M = 1 \& D = 1 \& Y = 0)) > P(E = 1 | do(M = 0 \& D = 1 \& Y = 0))$

(5) $P(E = 1 | do(M = 1 \& T = 1 \& D = 1 \& Y = 0)) > P(E = 1 | do(M = 0 \& D = 1 \& Y = 0))$

Inequality (4) indicates that my action raises the probability of the explosion *under the contingency* – i.e. *holding fixed* – that (you make your decision but) don't place your Th-228 near the Geiger. The existence of this *contingent* probability-raising reflects the fact that there is a path – $\langle M, T, E \rangle$ – along which $M = 1$ promotes $E = 1$ (because $M = 1$ raises the probability of $E = 1$ when we hold fixed the values of all variables off that path). It is the existence of

such a path – representing the process via which $M = 1$ produces $E = 1$ – that appears to drive our intuitions about actual causation in this case (cf. Hitchcock 2001).

Inequality (5) indicates that, again holding fixed $D = 1$ and $Y = 0$, the probability of $E = 1$ is higher if I place my U-232 near the Geiger *and the threshold reading is reached* than if I'd simply never placed my U-232 near the Geiger in the first place. As will be seen, this requirement ensures that, not only is there a potential process via which $M = 1$ threatens to bring about $E = 1$, but that process is complete.

Since **AC(M-R)** implies that $M = 1$ is an actual cause of $E = 1$ relative to $\mathscr{M}_{Pre}$, **AC(S)** yields the (correct) result that $M = 1$ is an actual cause (simpliciter) of $E = 1$ provided that $\mathscr{M}_{Pre}$ is appropriate. $\mathscr{M}_{Pre}$ *is* appropriate. Clearly it satisfies (Partition) and (Independence). It satisfies (Naturalness) because all of the states that its variables represent are reasonably natural. It was stipulated that the $do(\cdot)$ function associated with $\mathscr{M}_{Pre}$ is such that (Veridicality) is satisfied. $\mathscr{M}_{Pre}$ does not represent the sort of 'non-serious' possibility that (Serious Possibilities) is introduced to rule out (cf. Hitchcock 2001; Woodward 2005, 86–91).

Finally, (Stability) is satisfied because the causal process from my action to the explosion is complete. Holding fixed $Y = 0$, the probability of the explosion if $M = 1$ *and* part(s) of this process occur(s) is higher than the probability of the explosion if simply $M = 0$. Any variable (whose values represent reasonably natural states, form a partition, and are logically and metaphysically independent from the variables in $\mathscr{V}_{Pre}$) that might be added to $\mathscr{V}_{Pre}$ either represents part of this process or it doesn't. If it does, its actual value represents *the occurrence* of part of the process. So, if it is added to $\mathscr{V}_{Pre}$, including it in $\mathbf{Z}$ will not prevent (**IN**) from holding for all subsets $\mathbf{Z}'$ of $\mathbf{Z}$. If it doesn't, then adding it to $\mathscr{V}_{Pre}$, including it in $\mathbf{W}$, and holding it fixed at its actual value as part of the assignment $\mathbf{W} = \mathbf{w}'$ will not make a difference to the fact that (**IN**) holds for all subsets $\mathbf{Z}'$ of $\mathbf{Z}$, since holding fixed $Y = 0$ as part of $\mathbf{W} = \mathbf{w}'$ is already sufficient to ensure this.

Probabilistic Actual Causation

AC(M-R) gives the verdict that $D = 1$ is *not* an actual cause of $E = 1$ relative to $\mathscr{M}_{Pre}$. Consider the partition of $\mathscr{V}_{Pre}\backslash D, E$ such that $\mathbf{W} = \{M\}$ and $\mathbf{Z} = \{Y, T\}$. Observe that:

(6) $$P(E = 1|do(D = 1\&M = 0)) > P(E = 1|do(D = 0\&M = 0))$$

And:

(7) $$P(E = 1|do(D = 1\&M = 1)) > P(E = 1|do(D = 0\&M = 1))$$

Thus, whichever possible value we hold fixed M at, the probability of $E = 1$ is higher if $D = 1$ than if $D = 0$. So $D = 1$ contingently raises the probability of $E = 1$.[12] That's because there's a path – $\langle D, Y, E\rangle$ – along which $D = 1$ promotes $E = 1$.

AC(M-R) nevertheless entails that $D = 1$ is *not* an actual cause of $E = 1$ relative to $\mathscr{M}_{Pre}$. Consider the subset $\{Y\}$ of $\mathbf{Z}$, and observe that:

(8) $$P(E = 1|do(D = 1\&Y = 0\&M = 0)) \leq P(E = 1|do(D = 0\&M = 0))$$

And:

(9) $$P(E = 1|do(D = 1\&Y = 0\&M = 1)) \leq P(E = 1|do(D = 0\&M = 1))$$

That is, whichever possible value we hold fixed M at, the probability of the explosion is no higher if you make your decision *but don't place your Th-228 near the Geiger* than if you'd

[12] The obtaining of just one of (6) or (7) would suffice to show this.

never made that decision in the first place. Thus (**IN**) does not hold for every subset of **Z** for this partition of variables no matter what values we assign to the variables in **W**. This reflects the fact that, because you didn't place your Th-228 near the Geiger, there is no complete causal process by which your decision produces the explosion. Your non-placement of your Th-228 'neutralizes' the danger of your decision causing the explosion.

Is there an alternative partition (**W**, **Z**) of $\mathscr{V}_{Pre}$ and assignment $\mathbf{W} = \mathbf{w}'$ such that (**IN**) holds for all subsets $\mathbf{Z}'$ of **Z**? (There need only be *one* for **AC(M-R)** to be satisfied.) There isn't. Assigning Y to **W** instead of **Z** won't help, since the value of Y 'screens off' D from E. So, where $Y \in \mathbf{W}$, no assignment $\mathbf{W} = \mathbf{w}'$ will be such that, holding fixed $\mathbf{W} = \mathbf{w}'$, the probability of $E = 1$ is higher when $D = 1$ (and the variables in $\emptyset \subseteq \mathbf{Z}$ are set to their actual values) than when $D = 0$. So (**IN**) doesn't hold for all subsets $\mathbf{Z}'$ of **Z** for any such partition.

On the other hand, if we leave Y in **Z** and also assign M to **Z**, then there are no variables in **W** to hold fixed. Now consider the subset $\{Y\}$ of **Z**, and observe that:[13]

$$P(E = 1|do(D = 1 \& Y = 0)) \leq P(E = 1|do(D = 0)) \tag{10}$$

So, with M assigned to **Z** it remains the case that (**IN**) doesn't hold for all subsets of **Z**.

So there's no partition of $\mathscr{V}_{Pre} \backslash D, E$ such that (**IN**) is satisfied for all subsets of **Z** when we consider $D = 1$ as a putative cause of $E = 1$. **AC(M-R)** therefore doesn't count $D = 1$ as an actual cause of $E = 1$ relative to $\mathscr{M}_{Pre}$.

But for **AC(S)** to count $D = 1$ as an actual cause of $E = 1$ *simpliciter*, there need only be one appropriate model relative to which **AC(M-R)** counts $D = 1$ as an actual cause of $E = 1$. Is there such a model? There isn't. Suppose a candidate such model includes Y. Because D is only relevant to E because of its relevance to Y, the value of Y 'screens off' the value of D

[13] Note: the fact that $Y = 0$ *due to an intervention* doesn't make $M = 1$ more likely.

from that of E. This means that, if Y is included in **W** in the partition (**W**, **Z**) of the model's variable set and held fixed (either at 1 or 0) as part of the assignment $\mathbf{W} = \mathbf{w}'$, then (**IN**) won't be satisfied for the empty subset of **Z**. Alternatively, if Y is included in **Z** then, no matter what other variables are included in the model and assigned to **W**, (**IN**) won't be satisfied for the subset $\{Y\}$ of **Z**. Specifically, because $D = 1$ only threatens to bring about $E = 1$ because it threatens to bring about $Y = 1$, no matter what we hold fixed by inclusion on both sides of (**IN**), the probability of $E = 1$ is no higher if $D = 1$ *and* $Y = 0$ than if simply $D = 0$.

So **AC(M-R)** doesn't count $D = 1$ as an actual cause of $E = 1$ relative to any appropriate model with Y in its variable set. This means that any otherwise appropriate model relative to which $D = 1$ *is* an actual cause of $E = 1$ can be expanded to a model in which $D = 1$ *isn't* an actual cause of $E = 1$ simply by the addition of Y. Provided the expanded model is appropriate, the original model violates (Stability) and is inappropriate. So **AC(S)** will correctly not count $D = 1$ as an actual cause *simpliciter* of $E = 1$.

Since the values of Y form a partition and represent natural states of affairs, (Partition) and (Naturalness) will be satisfied by the expanded model if they were satisfied by the original model. With regard to (Veridicality), it should be noted that there are multiple ways of expanding the original model via the addition of Y, each associated with a different $do(\cdot)$ function from elements of $\mathscr{C}^*$ to probability distributions over $\mathscr{F}^*$ (where $\mathscr{C}^*$ and $\mathscr{F}^*$ are generated by the expanded variable set in the way described in Section 4). In looking for an apt expanded model, we just select the one with the $do(\cdot)$ function that returns the objective chances on $\mathscr{F}^*$ that *would* obtain as a result of interventions bringing about the various elements of $\mathscr{C}^*$. With regard to (Serious Possibilities) note that, given your decision, your placing *and* your not placing your Th-228 near the Geiger are both salient possibilities in

(ProbPre). So it doesn't seem that the expanded model could represent any non-serious possibilities if the original model doesn't. (Independence) is a little trickier. Might not the original model include a variable whose values are logically or metaphysically related to those of Y? Given that the variables in the original model are assumed to satisfy (Partition) it seems that any variable logically or metaphysically related to Y – e.g. Y', which takes value $Y' = 0$ if you don't place your Th-228 near the Geiger, $Y' = 1$ if you place it 2.5-5cm from the Geiger, and $Y' = 2$ if you place it 0-2.5cm from the Geiger – will also be such that its actual value neutralizes the threat of $D = 1$ bringing about $E = 1$, so that **AC(M-R)** is not satisfied in the original model. The exception to this would be if the original model included a variable that represents a gerrymandered states of affairs – e.g. Y'', which takes value $Y'' = 1$ if you place your Th-228 near the Geiger *or* Obama is US president, and $Y'' = 0$ otherwise – in which case the original model will violate (Naturalness).

7. Conclusion

Drawing upon PCMs, an account of PAC has been given that gives a correct treatment of probabilistic preemption on intuitive grounds. Traditional accounts of PAC misdiagnose this central test case (Menzies, 1989, 1996; Hitchcock 2004). Examination of whether PCMs can help tackle some of the other outstanding problems of PAC is warranted.

REFERENCES

Anscombe, E. (1971). *Causality and Determination*. Cambridge: CUP.

Blanchard, T. and J. Schaffer (2016). Cause without Default. In H. Beebee, C. Hitchcock, and H. Price (Eds.), *Making a Difference*. Oxford: OUP.

Goldszmidt, M. and J. Pearl (1992). Rank-Based Systems. In *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, San Mateo, CA, pp. 661–672. Morgan Kaufmann.

Halpern, J. Y. (2014). Appropriate Causal Models and Stability of Causation. In *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning*, Palo Alto, CA, pp. 198–207. AAAI Press.

Halpern, J. Y. and C. Hitchcock (2010). Actual Causation and the Art of Modeling. In R. Dechter, H. Geffner, and J. Y. Halpern (Eds.), *Heuristics, Probability and Causality*, pp. 383–406. London: College Publications.

Halpern, J. Y. and J. Pearl (2001). Causes and Explanations: A Structural-Model Approach. Part I: Causes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, pp. 194–202. Morgan Kaufmann.

Halpern, J. Y. and J. Pearl (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal for the Philosophy of Science 56*, 843–87.

Hitchcock, C. (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy 98*, 194–202.

Hitchcock, C. (2004). Do All and Only Causes Raise the Probabilities of Effects? In J. Collins, N. Hall, and L. Paul (Eds.), *Causation and Counterfactuals*, pp. 403–417. Cambridge, MA: MIT Press.

Hitchcock, C. (2007). Prevention, Preemption, and the Principle of Sufficient Reason. *Philosophical Review 116*, 495–532.

Ismael, J. (2009). Probability in Deterministic Physics. *Journal of Philosophy 106*, 89–108.

Lewis, D. (1986). *Philosophical Papers*, Volume 2. Oxford: Oxford University Press.

Loewer, B. (2001). Determinism and Chance. *Studies in History and Philosophy of Science Part B 32*, 609–620.

Menzies, P. (1989). Probabilistic Causation and Causal Processes: A Critique of Lewis. *Philosophy of Science 56*, 642–663.

Menzies, P. (1996). Probabilistic Causation and the Pre-emption Problem. *Mind 105*, 85–117.

Paul, L. (2000). Aspect Causation. *Journal of Philosophy 97*, 235–256.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (Second ed.). Cambridge: CUP.

Reichenbach, H. (1971). *The Direction of Time*. Mineola, NY: Dover.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Suppes, P. (1970). *A Probabilistic Theory of Causality, Acta Philosophica Fennica*. Amsterdam: North-Holland.

Weslake, B. (2016). A Partial Theory of Actual Causation. Forthcoming in *British Journal for the Philosophy of Science*.

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford: OUP.

PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association -212-

When Journal Editors Play Favorites*

Remco Heesen†

June 28, 2016

Abstract

Should editors of scientific journals practice triple-blind reviewing? I consider two arguments in favor of this claim. The first says that insofar as editors' decisions are affected by information they would not have had under triple-blind review, an injustice is committed against certain authors. I show that even well-meaning editors would commit this wrong and I endorse this argument.

The second argument says that insofar as editors' decisions are affected by information they would not have had under triple-blind review, it will negatively affect the quality of published papers. I distinguish between two kinds of biases that an editor might have. I show that one of them has a positive effect on quality and the other a negative one, and that the combined effect could be either positive or negative. Thus I do not endorse the second argument in general. However, I do endorse this argument for certain fields, for which I argue that the positive effect does not apply.

*Thanks to Kevin Zollman and Liam Bright for valuable comments and discussion. This work was partially supported by the National Science Foundation under grant SES 1254291.

†Department of Philosophy, Baker Hall 161, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA. Email: rheesen@cmu.edu

1 Introduction

Journal editors occupy an important position in the scientific landscape. By making the final decision on which papers get published in their journal and which papers do not, they have a significant influence on what work is given attention and what work is ignored in their field (Crane 1967).

In this paper I investigate the following question: should the editor be informed about the identity of the author when she is deciding whether to publish a particular paper? Under a single- or double-blind reviewing procedure, the editor has access to information about the author, whereas under a triple-blind reviewing procedure she does not. So in other words the question is: should journals practice triple-blind reviewing?

Two kinds of arguments have been given in favor of triple-blind reviewing. One focuses on the treatment of the author by the editor. On this kind of argument, revealing identity information to the editor will lead the editor to (partially) base her judgment on irrelevant information (such as the gender of the author, or whether or not the editor is friends with the author). This harms the author, and is thus bad.

The second kind of argument focuses on the effect on the journal and its readers. Again, the idea is that the editor will base her judgment on identity information if given the chance to do so. But now the further claim is that as a result the journal will accept worse papers. After all, if a decision to accept or reject a paper is influenced by the editor's biases, this suggests that a departure has been made from a putative "objectively correct" decision. This harms the readers of the journal, and is thus bad.

Here I provide a philosophical discussion of the reviewing procedure to assess these arguments. I distinguish between two different ways the editor's judgment may be affected if the author's identity is revealed to her. First, the editor may treat authors she knows differently from authors she does not know. Second, the editor may treat authors differently based on their membership of some group (e.g., gender bias). My discussion focuses on the

following three claims.

My first claim is that the first kind of differential treatment the editor may display (based on whether she knows a particular author) actually benefits rather than harms the readers of the journal. This benefit is the result of a reduction in editorial uncertainty about the quality of submitted papers when she knows their authors. I construct a model to show in a formally precise way how such a benefit might arise—surprisingly, no assumption that the scientists the editor knows are somehow "better scientists" is required—and I cite empirical evidence that such a benefit indeed does arise. However, this benefit only applies in certain fields. I argue that in other fields (in particular, mathematics and the humanities) no significant reduction of uncertainty—and hence no benefit to the readers—occurs (section 2).

My second claim is that either kind of differential treatment the editor may display (based on whether she knows authors or based on bias against certain groups) harms authors. I argue that any instance of such differential treatment constitutes an epistemic injustice in the sense of Fricker (2007) against the disadvantaged author. If the editor is to be (epistemically) just, she should prevent such differential treatment, which can be done through triple-blind reviewing. So I endorse an argument of the first of the two kinds I identified above: triple-blind reviewing is preferable because not doing so harms authors (section 3).

My third claim is that whether differential treatment also harms the journal and its readers depends on a number of factors. Differential treatment by the editor based on whether she knows a particular author may benefit readers, whereas differential treatment based on bias against certain groups may harm them. Whether there is an overall benefit or harm depends on the strength of the editor's bias, the relative sizes of the different groups, and other factors, as I illustrate using the model. As a result I do not in general endorse the second kind of argument, that triple-blind reviewing is preferable because readers of the journal are harmed otherwise. However, I do endorse

this argument for fields like mathematics and the humanities, where I claim that the benefits of differential treatment (based on uncertainty reduction) do not apply (section 4).

Note that, in considering the ethical and epistemic effects of triple-blind reviewing, a distinction is made between the effects on the author and the effects on the readers of the journal. This reflects a growing understanding that in order to study the social epistemology of science, what is good for an individual inquirer must be distinguished from what is good for the wider scientific community (Kitcher 1993, Strevens 2003, Mayo-Wilson et al. 2011).

Zollman (2009) has studied the effects of different editorial policies on the number of papers published and the selection criteria for publication, but he does not focus specifically on the editor's decisions and the uncertainty she faces. Economists have studied models in which editor decisions play an important role (Ellison 2002, Faria 2005, Besancenot et al. 2012), but they have not distinguished between papers written by scientists the editor knows and papers by scientists unknown to her, and neither have they been concerned with biases the editor may be subject to. And some other economists have done empirical work investigating the differences between papers with and without an author-editor connection (Laband and Piette 1994, Medoff 2003, Smith and Dombrowski 1998, more on this later), but they do not provide a model that can explain these differences. This paper thus fills a gap in the literature.

2 A Model of Editor Uncertainty

As I said in the introduction, journal editors have a certain measure of power in a scientific community because they decide which papers get published.[1] An editor could use this discretionary power to the benefit of her friends or

[1]Different journals may have different policies, such as one in which associate editors make the final decision for papers in their (sub)field. Here, I simply define "the editor" to be whomever makes the final decision whether to publish a particular paper.

colleagues, or to promote certain subfields or methodologies over others. This phenomenon has been called *editorial favoritism*. If anecdotal evidence is to be believed, this phenomenon is widespread. Some systematic evidence of favoritism exists as well. Bailey et al. (2008a,b) find that academics believe editorial favoritism to be fairly prevalent, with a nonnegligible percentage claiming to have perceived it firsthand. Laband (1985) and Piette and Ross (1992) find that, controlling for citation impact and various other factors, papers whose author has a connection to the journal editor are allocated more journal pages than papers by authors without such a connection.[2]

In this paper, I refer to the phenomenon that editors are more likely to accept papers from authors they know than papers from authors they do not know as *connection bias*.

Academics tend to disapprove of this behavior (Sherrell et al. 1989, Bailey et al. 2008a,b). In both of the studies by Bailey et al., in which subjects were asked to rate the seriousness of various potentially problematic behaviors by editors and reviewers, this disapproval was shown (using a factor analysis) to be part of a general and strong disapproval of "selfish or cliquish acts" in the peer review process. Thus it appears that the reason for the disapproval of editors publishing papers by their friends and colleagues is that it shows the editor acting on private interests, rather than displaying the disinterestedness that is the norm in science (Merton 1942).

On the other hand, if connection bias was a serious worry for authors, one would expect this to be a major consideration for them in choosing where to submit their papers (i.e., submit to journals where they know the editor), but Ziobrowski and Gibler (2000) find that this is not the case.[3]

[2]Here, page allocation is used as a proxy for journal editors' willingness to push the paper. The more obvious variable to use here would be whether or not the paper is accepted for publication. Unfortunately, there are no empirical studies which measure the influence of a relationship between the author and the editor on acceptance decisions directly. Presumably this is because information about rejected papers is usually not available in these kinds of studies.

[3]In particular, authors who know an editor and thus could expect to profit from con-

Moreover, despite working scientists' disapproval, there is some evidence that connection bias improves the overall quality of accepted papers (Laband and Piette 1994, Medoff 2003, Smith and Dombrowski 1998). Does that mean scientists are misguided in their disapproval?

As indicated in the introduction, I distinguish between the effects of editors' biases on the authors of scientific papers on the one hand, and the effects on the readers of scientific journals on the other hand. In this section, I use a formal model to show that these two can come apart: connection bias may negatively affect scientists as authors while positively affecting scientists as readers. Note that in this section I focus only on connection bias. Subsequent sections consider other biases.

Consider a simplified scientific community consisting of a set of scientists. Each scientist produces a paper and submits it to the community's only journal which has one editor.

Some papers are more suitable for publication than others. I assume that this suitability for publication can be measured on a single numerical scale. For convenience I call this the *quality* of the paper. However, I remain neutral on how this notion should be interpreted, e.g., as an objective measure of the epistemic value of the paper (which is perhaps an aggregate of multiple relevant criteria), or as the number of times the paper would be cited in future papers if it was published, or as the average subjective value each member of the scientific community would assign to it if they read it.[4]

nection bias would find knowing the editor and the composition of the editorial board more generally to be important factors in deciding where to submit, contrary to Ziobrowski and Gibler's evidence (these factors are ranked twelfth and sixteenth in importance in a list of sixteen factors that might influence the decision where to submit). Similarly, authors who do not know an editor would find a lack of (perceived) connection bias and the composition of the editorial board to be important factors, but these rank only seventh and twelfth in importance in Ziobrowski and Gibler's study. In a similar survey by Mackie (1998, chapter 4), twenty percent of authors indicated that knowing the editor and/or her preferences is an important consideration in deciding where to submit a paper.

[4]For more on potential difficulties with interpreting the notion of quality, see Bright (2015).

Crucially, the editor does not know the quality of the paper at the time it is submitted. The aim of this section is to show how uncertainty about quality can lead to connection bias. To make this point as starkly as possible, I assume that the editor cares only about quality, i.e., she makes an estimate of the quality of a paper and publishes those and only those papers whose quality estimate is high.

Let q_i be the quality of the paper submitted by scientist i. Since there is uncertainty about the quality, q_i is modeled as a random variable. Since some scientists are more likely to produce high quality papers than others, the mean μ_i of this random variable may be different for each scientist. I assume that quality follows a normal distribution with fixed variance: $q_i \mid \mu_i \sim N(\mu_i, \sigma^2_{qu})$.

The assumptions of normality and fixed variance are made primarily to keep the mathematics simple. Below I make similar assumptions on the distribution of average quality in the scientific community and the distribution of reviewers' estimates of the quality of a paper. I see no reason to expect the results I present below to be different when any of these assumptions are changed.

If the editor knows scientist i, she has some prior information on the average quality of scientist i's work. This is reflected in the model by assuming that the editor knows the value of μ_i. For scientists she does not know, the editor is uncertain about the average quality of their work. All she knows is the distribution of average quality in the larger scientific community, which I also assume to be normal: $\mu_i \sim N(\mu, \sigma^2_{sc})$.

Note that I assume the scientific community to be homogeneous: the scientific community is split in two groups (those known by the editor and those not known by the editor) but average paper quality follows the same distribution in both groups. If I assumed instead that scientists known by the editor write better papers on average the results would be qualitatively similar to those I present below. If scientists known by the editor write worse

papers on average this would affect my results. However, since most journal editors are relatively central figures in their field (Crane 1967), this would be an implausible assumption except perhaps in isolated cases.

The editor's prior beliefs about the quality of a paper submitted by some scientist i reflects this difference in information. If she knows the scientist she knows the value of μ_i, and so her prior is $\pi(q_i \mid \mu_i) \sim N(\mu_i, \sigma^2_{qu})$. If the editor does not know scientist i she only knows the distribution of μ_i, rather than its exact value. Integrating out the uncertainty over μ_i yields a prior $\pi(q_i) \sim N(\mu, \sigma^2_{qu} + \sigma^2_{sc})$ for the quality of scientist i's paper.

When the editor receives a paper she sends it out for review. In the context of this model, the main purpose of the reviewer's report is to provide an estimate of the quality of the paper. But, I assume, even after reading the paper its quality cannot be established with certainty. Thus the reviewer's estimate r_i of the quality q_i is again a random variable. I assume that the reviewer's report is unbiased, i.e., its mean is the actual quality q_i of the paper. Once again I use a normal distribution to reflect the uncertainty: $r_i \mid q_i \sim N(q_i, \sigma^2_{rv})$.[5]

The editor uses the information from the reviewer's report to update her beliefs about the quality of scientist i's paper. I assume that she does this by Bayes conditioning. Thus, her posterior beliefs about the quality of the paper are $\pi(q_i \mid r_i)$ if she does not know the author, and $\pi(q_i \mid r_i, \mu_i)$ if she does.

The posterior distributions are themselves normal distributions whose

[5]The reviewer's report could reflect the opinion of a single reviewer, or the averaged opinion of multiple reviewers. The editor could even act as a reviewer herself, in which case the report reflects her findings which she has to incorporate in her overall beliefs about the quality of the paper. The assumption I make in the text can be used to cover any of these scenarios, as long as a given journal is fairly consistent in the number of reviewers used. If the number of reviewers is frequently different for different papers (and in particular when this difference correlates with the existence or absence of a connection between editor and author) the assumption of a fixed variance in the reviewer's report is unrealistic because a report from multiple reviewers may be thought to give more accurate information (reducing the variance) than a report from a single reviewer.

mean is a weighted average of r_i and the prior mean, as given in proposition 1 (for a proof, see DeGroot 2004, section 9.5, or any other textbook that covers Bayesian statistics).

Proposition 1.

$$\pi(q_i \mid r_i) \sim N\left(\mu_i^U, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right),$$
$$\pi(q_i \mid r_i, \mu_i) \sim N\left(\mu_i^K, \frac{\sigma_{qu}^2\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right),$$

where

$$\mu_i^U = \frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\mu,$$
$$\mu_i^K = \frac{\sigma_{qu}^2}{\sigma_{qu}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\mu_i.$$

When does the editor choose to publish a paper? Here I assume that she publishes any paper whose posterior mean is above some threshold q^*. So a paper written by a scientist unknown to the editor is published if $\mu_i^U > q^*$ and a paper written by a scientist known to the editor is published if $\mu_i^K > q^*$. This corresponds to being at least 50% confident that the paper's quality is above the threshold. Other standards could be used (risk-averse standards might require more than 50% confidence that the paper is above some threshold, while risk-loving standards might require less; in these cases the threshold value needs to be adapted to keep the total number of accepted papers constant) but for my purposes here it does not much matter.

Now compare the probability that the paper of an arbitrary scientist i unknown to the editor is published to the probability that the paper of an arbitrary scientist known by the editor is published. For this purpose it is useful to determine the probability distribution of the posterior means (see appendix A for proofs of this and subsequent results).

Proposition 2. *The posterior means are normally distributed, with* $\mu_i^U \sim N(\mu, \sigma_U^2)$ *and* $\mu_i^K \sim N(\mu, \sigma_K^2)$. *Here,*

$$\sigma_U^2 = \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} \qquad \textit{and} \qquad \sigma_K^2 = \frac{\sigma_{qu}^4 + \sigma_{sc}^2(\sigma_{qu}^2 + \sigma_{rv}^2)}{\sigma_{qu}^2 + \sigma_{rv}^2}.$$

Moreover, if $\sigma_{sc}^2 > 0$ *and* $\sigma_{rv}^2 > 0$, *then* $\sigma_U^2 < \sigma_K^2$.

The main result of this section, which establishes the existence of connection bias in the model, is a consequence of proposition 2. It says that the editor is more likely to publish a paper written by an arbitrary author she knows than a paper written by an arbitrary author she does not know, whenever $q^* > \mu$ (for any positive value of σ_{sc}^2 and σ_{rv}^2). Since $q^* = \mu$ would mean that exactly half of all papers gets published, the condition amounts to a requirement that the journal's acceptance rate is less than 50%. This is true of most reputable journals in most fields (physics being a notable exception). When acceptance rates are above 50% editorial favoritism is also much less of a concern in the first place.

Theorem 3. *If* $q^* > \mu$, $\sigma_{sc}^2 > 0$, *and* $\sigma_{rv}^2 > 0$, *the acceptance probability for authors known to the editor is higher than the acceptance probability for authors unknown to the editor, i.e.,* $\Pr\left(\mu_i^K > q^*\right) > \Pr\left(\mu_i^U > q^*\right)$.

Theorem 3 shows that in the model I presented, any journal with an acceptance rate lower than 50% will be seen to display connection bias. Thus I have established the surprising result that an editor who cares only about the quality of the papers she publishes may end up publishing more papers by her friends and colleagues than by scientists unknown to her, even if her friends and colleagues are not, as a group, better scientists than average.

Why does this surprising result hold? The theorem follows immediately from proposition 2, which says that the distribution of μ_i^U is less "spread out" than the distribution of μ_i^K ($\sigma_U^2 < \sigma_K^2$). This happens because μ_i^U is a

weighted average of μ and r_i, keeping it relatively close to the overall mean μ compared to μ_i^K, which is a weighted average of μ_i and r_i (which tend to differ from μ in the same direction).

Because the editor treats papers by authors she knows differently from papers by authors she does not know, authors unknown to the editor are arguably harmed. I pick up this point in section 3 and argue that this constitutes an epistemic injustice against those authors.

What I have shown so far is that an editor who uses information about the average quality of papers produced by scientists she knows in her acceptance decisions will find that scientists she knows produce on average more papers that meet her quality threshold. This is a subjective statement: the editor believes that more papers by scientists she knows meet her threshold. Does this translate into an objective effect? That is, does the extra information the editor has available about scientists she knows allow her to publish better papers from them than from scientists she does not know?

In order to answer this question I need to compare the average quality of accepted papers. More formally, I want to compare the expected value of the quality of a paper, conditional on meeting the publication threshold, given that the author is either known to the editor or not.

Proposition 4. *If $\sigma_{sc}^2 > 0$, and $\sigma_{rv}^2 > 0$, the average quality of accepted papers from authors known to the editor is higher than the average quality of accepted papers from authors unknown to the editor, i.e., $\mathbb{E}[q_i \mid \mu_i^K > q^*] > \mathbb{E}[q_i \mid \mu_i^U > q^*]$.*

Proposition 4 shows that the editor can use the extra information she has about scientists she knows to improve the average quality of the papers published in her journal. In other words, the surprising result is that the editor's connection bias actually benefits rather than harms the readers of the journal. It is thus fair to say that, in the model, the editor can use her connections to "identify and capture high-quality papers", as Laband and

Piette (1994) suggest.[6]

To what extent does this show that the connection bias observed in reality is the result of editors capturing high-quality papers, as opposed to editors using their position of power to help their friends? At this point the model is seen to yield an empirical prediction. If connection bias is (primarily) due to capturing high-quality papers, the quality of papers by authors the editor knows should be higher than average, as shown in the model. If, on the other hand, connection bias is (primarily) a result of the editor accepting for publication papers written by authors she knows even though they do not meet the quality standards of the journal, then the quality of papers by authors the editor knows should (presumably) be lower than average.

If subsequent citations are a good indication of the quality of a paper,[7] a simple regression can test whether accepted papers written by authors with an author-editor connection have a higher or a lower average quality than papers without such a connection. This empirical test has been carried out a number of times, and the results univocally favor the hypothesis that editors use their connections to improve the quality of published papers (Laband and Piette 1994, Smith and Dombrowski 1998, Medoff 2003).

Note that in the above results, nothing depends on the sizes of the variances σ^2_{qu}, σ^2_{sc}, and σ^2_{rv}. This is because these results are qualitative. The variances do matter when the acceptance rate and average quality of papers are compared quantitatively. For example, reducing σ^2_{rv} (making the reviewer's report more accurate) makes the differences in the acceptance rate and average quality of papers smaller.

[6]This result applies to connection bias only. Below I consider other biases the editor might have, which yields more nuanced conclusions.

[7]Recall that I have remained neutral on how the notion of quality should be interpreted. If quality is simply defined as "the number of citations this paper would get if it were published" the connection between quality and citations is obvious. Even on other interpretations of quality, citations have frequently been viewed as a good proxy measure (Cole and Cole 1967, 1968, Medoff 2003). This practice has been defended by Cole and Cole (1971) and Clark (1957, chapter 3), and criticized by Lindsey (1989) and Heesen (forthcoming).

Note also that the results depend on the assumption that σ^2_{sc} and σ^2_{rv} are positive. What is the significance of these assumptions?

If $\sigma^2_{rv} = 0$, i.e., if there is no variance in the reviewer's report, the reviewer's report describes the quality of the paper with perfect accuracy. In this case the "extra information" the editor has about authors she knows is not needed, and so there is no difference in acceptance rate or average quality based on whether the editor knows the author. But it seems unrealistic to expect reviewer's reports to be this accurate.

If $\sigma^2_{sc} = 0$ there is either no difference in the average quality of papers produced by different authors, or learning the identity of the author does not tell the editor anything about the expected quality of that scientist's work. In this case there is no value to the editor (with regard to determining the quality of the submitted paper) in learning the identity of the author. So here also there is no difference in acceptance rate or average quality based on whether the editor knows the author.

Under what circumstances should the identity of the author be expected to tell the editor something useful about the quality of a submitted paper? This seems to be most obviously the case in the lab sciences. The identity of the author, and hence the lab at which the experiments were performed, can increase or decrease the editor's confidence that the experiments were performed correctly, including all the little checks and details that are impossible to describe in such a paper. In a scientific paper, "[a]s long as the conclusions depend at least in part on the results of some experiment, the reader must rely on the author's (and perhaps referee's) testimony that the author really performed the experiment exactly as claimed, and that it worked out as reported" (Easwaran 2009, p. 359).

But in other fields, in particular mathematics and some or all of the humanities, there is no need to rely on the author's reputation. This is because in these fields the paper itself is the contribution, so it is possible to judge papers in isolation of how or by whom they were created. Easwaran

(2009) discusses this in detail for mathematics, and briefly (in his section 4) for philosophy. And in fact there exists a norm that this is how they should be judged: "Papers will rely only on premises that the competent reader can be assumed to antecedently believe, and only make inferences that the competent reader would be expected to accept on her own consideration." (Easwaran 2009, p. 354).

Arguably then, the advantage (see theorem 3 and proposition 4) conferred by revealing identity information about the author to the editor applies only in certain fields. The relevant fields are those where part of the information in the paper is conferred on the authority of testimony, in particular those where experimental results are reported. Even in those fields, of course, what is being testified is supposed to be reproducible by the reader. But this is still different from the case in mathematics and the humanities, where a careful reading of a paper itself constitutes a reproduction of its argument. In these latter fields there is no relevant information to be learned from the identity of the author (i.e., $\sigma^2_{sc} = 0$), or, at least, the publishing norms in these fields suggest that their members believe this to be the case.

3 Bias As an Epistemic Injustice

The previous section discussed a formal model of editorial uncertainty about paper quality. The first main result, theorem 3, established the existence of connection bias in this model: authors known by the editor are more likely to see their paper accepted than authors unknown to the editor. The second main result, proposition 4, showed that connection bias benefits the readers of the journal by improving the average quality of accepted papers.

Despite the benefit to the readers, I claim that authors are harmed by connection bias. In this section I argue that an instance of connection bias constitutes an *epistemic injustice* in the sense of Fricker (2007). Then I argue that the editor is likely to display other biases as well, and that instances of

these also constitute epistemic injustices.

The type of epistemic justice that is relevant here is *testimonial injustice*. Fricker (2007, pp. 17–23) defines a testimonial injustice as a case where a speaker suffers a credibility deficit for which the hearer is ethically and epistemically culpable, rather than being due to innocent error.

Testimonial injustices may arise in various ways. Fricker is particularly interested in what she calls "the central case of testimonial injustice" (Fricker 2007, p. 28). This kind of injustice results from a *negative identity-prejudicial stereotype*, which is defined as follows:

> A widely held disparaging association between a social group and one or more attributes, where this association embodies a generalization that displays some (typically, epistemically culpable) resistance to counter-evidence owing to an ethically bad affective investment. (Fricker 2007, p. 35)

Because the stereotype is widely held, it produces *systematic* testimonial injustice: the relevant social group will suffer a credibility deficit in many different social spheres.

Applying this to the phenomenon of connection bias, it is clear that this is not an instance of the central case of testimonial injustice. This would entail that there is some negative stereotype associated with scientists unknown to the editor, as a group, which is not normally the case. So I set the central case aside (I return to it below) and focus on the question whether connection bias can produce (non-central cases of) testimonial injustice.

Suppose scientist i and scientist i' tend to produce papers of the same quality, which is above average in the population ($\mu_i = \mu_{i'} > \mu$). Suppose further that the actual papers they have produced on this occasion are of the same quality ($q_i = q_{i'}$) and have received similar reviewer reports ($r_i = r_{i'}$). If scientist i is not known to the editor, but scientist i' is, then the paper

written by scientist i' is likely to be evaluated more highly by the editor.[8] If the publication threshold q^* is somewhere in between the two evaluations then only scientist i' will have her paper accepted.

In this example, the scientists produced papers of equal quality that were evaluated differently. So scientist i suffers a credibility deficit. This deficit is not due to innocent error, as it would be if, e.g., random variation led to different reviewer reports (i.e., $r_i < r_{i'}$). The deficit is also not due to the editor's use of generally reliable information about the two scientists, as it would be if there was a genuine difference in the average quality of the papers they produce (i.e., $\mu_i < \mu_{i'}$).

Is this credibility deficit suffered by scientist i ethically and epistemically culpable on the part of the editor? On the one hand, as I stressed in section 2, the editor is simply making maximal use of the information available to her. It just so happens that she has more information about scientists she knows than about others. But that is hardly the editor's fault: she cannot be expected to know everyone's work. Is it incumbent upon her to get to know the work of every scientist who submits a paper?

This may well be too much to ask. But an alternative option is to remove all information about the authors of submitted papers. This can be done by using a triple-blind reviewing procedure, in which the editor does not know the identity of the author, and hence is prevented from using information about scientists she knows in her evaluation. Using such a procedure, at least all scientists are treated equally: any scientist who writes a paper of a given quality has the same chance of seeing that paper accepted.

So a credibility deficit occurs which harms scientist i: her paper is rejected. Moreover, it harms her specifically as an epistemic agent: the rejection of the paper reflects a judgment of the quality of her scientific work. And

[8] The editor's posterior mean for the quality of scientist i's paper is μ_i^U and her posterior mean for scientist i''s paper is $\mu_{i'}^K = \mu_i^K$, with $\mu_i^U < \mu_{i'}^K$ whenever $\sigma_{sc}^2(r_i - \mu_i) < (\sigma_{qu}^2 + \sigma_{rv}^2)(\mu_i - \mu)$. The claim in the text is then justified by the fact that $\Pr(\sigma_{sc}^2(r_i - \mu_i) < (\sigma_{qu}^2 + \sigma_{rv}^2)(\mu_i - \mu) \mid \mu_i > \mu) > 1/2$, assuming $\sigma_{sc}^2 > 0$ and $\sigma_{rv}^2 > 0$.

this harm could have been prevented by the editor by using a triple-blind reviewing procedure.

I conclude that the editor is ethically and epistemically culpable for this credibility deficit, and hence a testimonial injustice is committed against scientist i. However, one may insist that it cannot be the case that the editor is committing a wrong simply in virtue of using relevant information that is available to her. An evidentialist in particular may say that it cannot possibly be an epistemic wrong to take into account all relevant information.

I disagree, for the reasons just given, but I need not insist on this point. Even if it is granted that the editor does not commit an injustice by using the information that is available to her, the end result is still that scientist i is harmed as an epistemic agent. She has produced a paper of equal quality to scientist i''s, and yet it is not published.

Moreover, the presence of scientist i' is irrelevant. Any time a paper from an author unknown to the editor is rejected which would have been accepted had the editor known the author (all else being equal), that author is harmed. So even if one insists that differential editorial treatment resulting from connection bias is not culpable on the part of the editor, connection bias still harms authors whenever it influences acceptance decisions.

In the model of section 2, and the above discussion, I assumed that connection bias is the only bias journal editors display. The literature on implicit bias suggests that this is not true. For example, "[i]f submissions are not anonymous to the editor, then the evidence suggests that women's work will probably be judged more negatively than men's work of the same quality" (Saul 2013, p. 45). Evidence for this claim is given by Wennerås and Wold (1997), Valian (1999, chapter 11), Steinpreis et al. (1999), Budden et al. (2008), and Moss-Racusin et al. (2012).[9] So women scientists are at

[9]These citations show that the work of women in academia is undervalued in various ways. None of them focus specifically on editor evaluations, but they support Saul's claim unless it is assumed that journal editors as a group are significantly less biased than other academics.

a disadvantage simply because of their gender identity. Similar biases exist based on other irrelevant aspects of scientists' identity, such as race or sexual orientation (see Lee et al. 2013, for a critical survey of various biases in the peer review system). As Crandall (1982, p. 208) puts it: "The editorial process has tended to be run as an informal, old-boy network which has excluded minorities, women, younger researchers, and those from lower-prestige institutions".

I use *identity bias* to refer to these kinds of biases. Any time a paper is rejected because of identity bias (i.e., the paper would have been accepted if the relevant part of the author's identity had been different, all else being equal), a testimonial injustice occurs for the same reasons outlined above. Moreover, here the editor is culpable for having these biases.

Unlike instances resulting from connection bias, testimonial injustices resulting from identity bias can be instances of the central case of testimonial injustice, in which the credibility deficit results from a negative identity-prejudicial stereotype. The evidence suggests that negative identity-prejudicial stereotypes affect the way people (not just men) judge women's work, even when the person judging does not consciously believe in these stereotypes. Moreover, those who think highly of their ability to judge work objectively and/or are primed with objectivity are affected more rather than less (Uhlmann and Cohen 2007, Stewart and Payne 2008, p. 1333). Similar claims plausibly hold for biases based on race or sexual orientation. Biases based on academic affiliation are not usually due to negative identity-prejudicial stereotypes, as these do not generally affect other aspects of the scientist's life.

So both connection bias and identity bias are responsible for injustices against authors. This is one way to spell out the claim that authors are harmed when journal editors do not use a triple-blind reviewing procedure. This constitutes the first kind of argument for triple-blind reviewing which I mentioned in the introduction, and which I endorse based on these consid-

erations.

4 The Effect of Bias on Quality

The second kind of argument I mentioned in the introduction claims that failing to use triple-blind reviewing harms the journal and its readers, because it would lower the average quality of accepted papers. In section 2 I argued that connection bias actually has the opposite effect: it increases average quality. In this section I complicate the model to include identity bias.

Recall that the editor displays identity bias if she is more or less likely to publish papers from a certain group of scientists based on some aspect of their identity, e.g., their gender. I incorporate this in the model by assuming the editor consistently undervalues members of one group (and overvalues the others). More precisely, she believes the average quality of papers produced by any scientist i from the group she is biased against to be lower than it really is by some constant quantity ε. Conversely, the average quality of papers written by any scientist not belonging to this group is raised by δ.[10] So the editor has a different prior for the two groups; I use π_A to denote her prior for the quality of papers written by scientists she is biased against, and π_F for her prior for scientists she is biased in favor of.

As before, the editor may be familiar with a given scientist's work (i.e., she knows the average quality of that scientist's papers) or not. So there are now four groups. If scientist i is known to the editor and belongs to the stigmatized group the editor's prior distribution on the quality of scientist i's paper is $\pi_A(q_i \mid \mu_i) \sim N(\mu_i - \varepsilon, \sigma^2_{qu})$. If scientist i is known to the editor but is not in the stigmatized group the prior is $\pi_F(q_i \mid \mu_i) \sim N(\mu_i + \delta, \sigma^2_{qu})$. If

[10]This is a simplifying assumption: one could imagine having biases against multiple groups of different strengths, or biases whose strength has some random variation, or biases which intersect in various ways (Collins and Chepp 2013, Bright et al. 2016). However, the assumption in the main text suffices to make the point I want to make. It should be fairly straightforward to extend my results to more complicated cases like the ones just described.

scientist i is not known to the editor and is in the stigmatized group the prior is $\pi_A(q_i) \sim N(\mu - \varepsilon, \sigma_{qu}^2 + \sigma_{sc}^2)$. And if scientist i is not known to the editor and not in the stigmatized group the prior is $\pi_F(q_i) \sim N(\mu + \delta, \sigma_{qu}^2 + \sigma_{sc}^2)$.[11]

The next few steps in the development are analogous to that in section 2. After the reviewer's report comes in the editor updates her beliefs about the quality of the paper, yielding the following posterior distributions.

Proposition 5.

$$\pi_A(q_i \mid r_i, \mu_i) \sim N\left(\mu_i^{KA}, \frac{\sigma_{qu}^2 \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right),$$
$$\pi_F(q_i \mid r_i, \mu_i) \sim N\left(\mu_i^{KF}, \frac{\sigma_{qu}^2 \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right),$$
$$\pi_A(q_i \mid r_i) \sim N\left(\mu_i^{UA}, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right),$$
$$\pi_F(q_i \mid r_i) \sim N\left(\mu_i^{UF}, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right),$$

where

$$\mu_i^{KA} = \mu_i^K - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \qquad \mu_i^{KF} = \mu_i^K + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2},$$
$$\mu_i^{UA} = \mu_i^U - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \qquad \mu_i^{UF} = \mu_i^U + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}.$$

As before, the paper is published if the posterior mean (μ_i^{KA}, μ_i^{KF}, μ_i^{UA}, or μ_i^{UF}) exceeds the threshold q^*. The respective distributions of the posterior

[11] Note that I assume that the editor displays bias against scientists in the stigmatized group regardless of whether she knows them or not. Under a reviewing procedure that is not triple-blind, the editor learns at least the name and affiliation of any scientist who submits a paper. This information is usually sufficient to determine with reasonable certainty the scientist's gender. So at least for gender bias it seems reasonable to expect the editor to display bias even against scientists she does not know. Conversely, because negative identity-prejudicial stereotypes can work unconsciously, it does not seem reasonable to expect that the editor can withhold her bias from scientists she knows.

means determine how likely this is. These distributions are given in the next proposition.

Proposition 6. *The posterior means are normally distributed, with*

$$\begin{aligned}
\mu_i^{KA} &\sim N\left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2\right), \\
\mu_i^{KF} &\sim N\left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2\right), \\
\mu_i^{UA} &\sim N\left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2\right), \\
\mu_i^{UF} &\sim N\left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2\right).
\end{aligned}$$

This yields the within-group acceptance rates and the unsurprising result that the editor is less likely to publish papers by scientists she is biased against.

Theorem 7. *If $\varepsilon > 0$, $\delta > 0$, and $\sigma_{rv}^2 > 0$, the acceptance probability for authors the editor is biased against is lower than the acceptance probability for authors the editor is biased in favor of (keeping fixed whether or not the editor knows the author). That is,*

$$\Pr\left(\mu_i^{KA} > q^*\right) < \Pr\left(\mu_i^{KF} > q^*\right) \text{ and } \Pr\left(\mu_i^{UA} > q^*\right) < \Pr\left(\mu_i^{UF} > q^*\right).$$

Theorem 7 establishes the existence of identity bias in the model: authors that are subject to a negative identity-prejudicial stereotype are less likely to see their paper accepted than authors who are not. As I argued in section 3, whenever a paper is rejected due to identity bias this constitutes a testimonial injustice against the author.

Now I turn my attention to the effect that identity bias has on the average quality of accepted papers. In the current version of the model there is both

connection bias and identity bias. Connection bias has been shown to have a positive effect on average quality (see section 2). Whether the net effect of connection bias and identity bias is positive or negative depends on various parameters, as I illustrate below.

The benchmark for judging the average quality of accepted papers under a procedure subject to connection bias and identity bias is a *triple-blind reviewing procedure* under which the editor is not informed of the identity of the scientist. As a result, she is both unable to use information about the average quality of a given scientist's papers and unable to display bias against scientists based on their identity.

Under this triple-blind procedure, the editor's prior distribution for the quality of any submitted paper is $\pi(q_i) \sim N(\mu, \sigma_{qu}^2 + \sigma_{sc}^2)$, i.e., the prior I used in section 2 when the author was unknown to the editor. Hence, under this procedure, the posterior is $\pi(q_i \mid r_i)$, the posterior mean is $\mu_i^U \sim N(\mu, \sigma_U^2)$, the probability of acceptance is $\Pr(\mu_i^U > q^*)$ and the average quality of accepted papers is $\mathbb{E}[q_i \mid \mu_i^U > q^*]$.

In contrast, I refer to the reviewing procedure that is subject to connection bias and identity bias as the *non-blind procedure*. The overall probability that a paper is accepted under the non-blind procedure depends on the relative sizes of the four groups. I use p_{KA} to denote the fraction of scientists known to the editor that she is biased against, p_{KF} for the fraction known to the editor that she is biased in favor of, p_{UA} for unknown scientists biased against, and p_{UF} for unknown scientists biased in favor of. These fractions are nonnegative and sum to one.

Let A_i denote the event that scientist i's paper is accepted under the non-blind procedure. The overall probability of acceptance under this procedure is

$$\begin{aligned}\Pr\left(A_i\right) = p_{KA}\Pr\left(\mu_i^{KA} > q^*\right) + p_{KF}\Pr\left(\mu_i^{KF} > q^*\right) \\ + p_{UA}\Pr\left(\mu_i^{UA} > q^*\right) + p_{UF}\Pr\left(\mu_i^{UF} > q^*\right).\end{aligned}$$

The average quality of accepted papers can then be written as $\mathbb{E}[q_i \mid A_i]$. I want to compare $\mathbb{E}[q_i \mid A_i]$ to $\mathbb{E}[q_i \mid \mu_i^U > q^*]$, the average quality of accepted papers under a triple-blind procedure.[12]

In the remainder of this section I assume that the editor's biases are such that she believes the average quality of all submitted papers to be equal to μ. In other words, her bias against the stigmatized group is canceled out on average by her bias in favor of those not in the stigmatized group, weighted by the relative sizes of those groups:

$$\left(p_{KA} + p_{UA}\right)\varepsilon = \left(p_{KF} + p_{UF}\right)\delta.$$

I use the above equation to fix the value of δ, reducing the number of free parameters by one. The equation amounts to a kind of commensurability requirement for the two procedures because it guarantees that the editor perceives the average quality of submitted papers to be the same regardless of whether or not a triple-blind procedure is used.

As far as I can tell there are no interesting general conditions on the parameter values that determine whether the non-blind procedure or the triple-blind procedure will lead to a higher average quality of accepted papers. The question I will explore now, using some numerical examples, is how biased the editor needs to be for the epistemic costs of her identity bias to outweigh the epistemic benefits resulting from connection bias.

In order to generate numerical data values have to be chosen for the

[12]Expressions for $\Pr(A_i)$ and $\mathbb{E}[q_i \mid A_i]$ using only the parameter values and standard functions are given in lemma 11 in appendix A. These expressions are used to generate the numerical results below.

parameters. First I set $\mu = 0$ and $q^* = 2$. Since quality is an interval scale in this model, these choices are arbitrary. For the variances σ^2_{qu}, σ^2_{sc}, and σ^2_{rv}, I choose a "small" and a "large" value (1 and 4 respectively).

For the sizes of the four groups, I assume that there is no correlation between whether the editor knows an author and whether the editor has a bias against that author (so, e.g., the percentage of women among scientists the editor knows is equal to the percentage of women among scientists the editor does not know). I consider two cases for the editor's identity bias: either she is biased against half the set of authors (and so biased in favor of the other half) or the group she is biased against is a 30 % minority.[13] Similarly, I consider the case in which the editor knows half of all scientists submitting papers, and the case in which the editor knows 30 % of them.

As a result, there are 32 possible settings of the parameters (2^3 choices for the variances times 2^2 choices for the group sizes). Whether the triple-blind procedure or the non-blind procedure is epistemically preferable depends on the value of ε (and the value of δ determined thereby).

It follows from proposition 4 that when $\varepsilon = 0$ the non-blind procedure helps rather than harms the readers of the journal by increasing average quality relative to the triple-blind procedure. If ε is positive but relatively small, this remains true, but when ε is relatively big, the non-blind procedure harms the readers. This is because the average quality of published papers under the non-blind procedure decreases continuously as ε increases (I do not prove this, but it is easily checked for the 32 cases I consider).

The interesting question, then, is where the turning point lies. How big does the editor's bias need to be in order for the negative effects of identity bias on quality to cancel out the positive effects of connection bias?

[13]Bruner and O'Connor (forthcoming) note that certain dynamics in academic life can lead to identity bias against groups as a result of the mere fact that they are a minority. Here I consider both the case where the stigmatized group is a minority (and is possibly stigmatized as a result of being a minority, as Bruner and O'Connor suggest) and the case where it is not (and so presumably the negative identity-prejudicial stereotype has some other source).

I determine the value of ε for which the average quality of published papers under the non-blind procedure and the triple-blind procedure is the same for each of the 32 cases. But reporting these numbers directly does not seem particularly useful, as ε is measured in "quality points" which do not have a clear interpretation outside of the model.

To give a more meaningful interpretation of these values of ε as measuring "size of bias", I calculate the average rate of acceptance of papers from authors the editor is biased against and the average rate of acceptance of papers from authors the editor is biased in favor of.[14] The difference between these numbers gives an indication of the size of the editor's bias: it measures (in percentage points, abbreviated pp) how many more papers the editor accepts from authors she is biased in favor of, compared to those she is biased against.

This difference is reported for the 32 cases in figure 1. To provide a sense of scale for these numbers, I plot them against the acceptance rate that the triple-blind procedure would have for those values of the parameters, i.e., $\Pr(\mu_i^U > q^*)$.

Already with this small sample of 32 cases, a large variation of results can be observed. I illustrate this by looking at two cases in detail.

First, suppose that $\sigma_{qu}^2 = \sigma_{sc}^2 = 1$ and $\sigma_{rv}^2 = 4$. In this extreme case the triple-blind procedure has an acceptance rate as low as 0.72 %. If the groups are all of equal size ($p_{KA} = p_{KF} = p_{UA} = p_{UF} = 1/4$) then under the non-blind procedure the acceptance rate for authors the editor is biased in favor of needs to be as much as 2.66 pp higher than the acceptance rate for authors the editor is biased against, in order for the average quality under

[14] These are calculated without regard for whether the editor knows the author or not. In particular, the rate of acceptance for authors the editor is biased against is

$$\frac{p_{KA}\Pr\left(\mu_i^{KA} > q^*\right) + p_{UA}\Pr\left(\mu_i^{UA} > q^*\right)}{p_{KA} + p_{UA}}, \text{ and } \frac{p_{KF}\Pr\left(\mu_i^{KF} > q^*\right) + p_{UF}\Pr\left(\mu_i^{UF} > q^*\right)}{p_{KF} + p_{UF}}$$

is the rate of acceptance for authors the editor is biased in favor of.

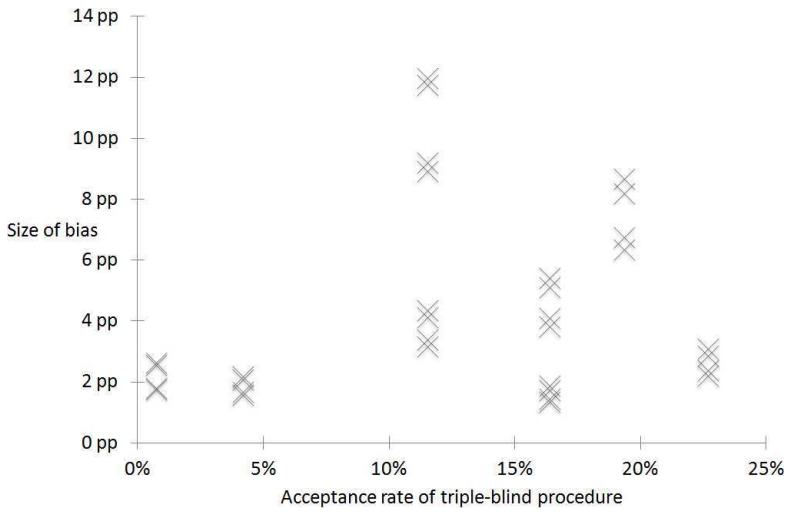


Figure 1: The minimum size of the editor's bias such that the quality costs of the non-blind procedure outweigh its benefits (given as a percentage point difference in acceptance rates), in 32 cases, plotted as a function of the acceptance rate of the corresponding triple-blind procedure.

the two procedures to be equal. Clearly a 2.66 pp bias is very large for a journal that only accepts less than 1 % of papers. If the bias is any less than that there is no harm to the readers in using the non-blind procedure.

Second, suppose that $\sigma^2_{qu} = \sigma^2_{sc} = 4$ and $\sigma^2_{rv} = 1$. Then the triple-blind procedure has an acceptance rate of 22.66 %. If, moreover, the editor knows relatively few authors ($p_{KA} = p_{KF} = 0.15$, $p_{UA} = p_{UF} = 0.35$) then the acceptance rate for authors the editor is biased in favor of needs to be only 2.23 pp higher than the acceptance rate for authors the editor is biased against, in order for the quality costs of the non-blind procedure to outweigh its benefits. For a journal accepting about 23 % of papers that means that even if the identity bias of the editor is relatively mild the journal's readers are harmed if the non-blind procedure is used.

Based on these results, and the fact that the parameter values are unlikely to be known in practice, it is unclear whether the non-blind procedure

or the triple-blind procedure will lead to a higher average quality of published papers for any particular journal.[15] So in general it is not clear that an argument that the non-blind procedure harms the journal's readers can be made. At the same time, a general argument that the non-blind procedure helps the readers is not available either. Given this, I am inclined to recommend a triple-blind procedure for all journals because not doing so harms the authors.

If there was reason to believe that the editor's bias was very small, there might be a case for the non-blind procedure using considerations of average quality. Based on the empirical evidence I cited in section 3, it seems unlikely that any editor could make such a case convincingly today. But if identity bias were someday to be eliminated or severely mitigated, this question may be worth revisiting.

So far I have argued in this section that in the presence of the positive effect of connection bias on quality, the net effect of connection bias and identity bias on quality is unclear. But I argued in section 2 that the positive effect of connection bias may only exist in certain fields. In fields where papers rely partially on the author's testimony there is value in knowing the identity of the author. But in other fields such as mathematics and some of the humanities testimony is not taken to play a role—the paper itself constitutes the contribution to the field—and so arguably there is no value in knowing the identity of the author.

In those fields, then, there is no quality benefit from connection bias, but there is still a quality cost from identity bias. So here the strongest case for the triple-blind procedure emerges, as the non-blind procedure harms both authors and readers.

[15]Note that the evidence collected by Laband and Piette (1994) does not help settle this question, as they do not directly compare the triple-blind and the non-blind procedure. Their evidence supports a positive epistemic effect of connection bias, but not a verdict on the overall epistemic effect of triple-blinding.

5 Conclusion

In this paper I have considered two types of arguments for triple-blind review: one based on the consequences for the author and one based on the consequences for the readers of the journal.

I have argued that the non-blind procedure introduces differential treatment of scientific authors. In particular, editors are more likely to publish papers by authors they know (connection bias, theorem 3) and less likely to publish papers by authors they apply negative identity-prejudicial stereotypes to (identity bias, theorem 7). Whenever a paper is rejected as a result of one of these biases an epistemic injustice (in the sense of Fricker 2007) is committed against the author. This is an argument in favor of triple-blinding based on consequences for the author.

From the readers' perspective the story is more mixed. Generally speaking connection bias has a positive effect on the quality of published papers and identity bias a negative one. Thus whether the readers are better off under the triple-blind procedure depends on how exactly these effects trade off, which is highly context-dependent, or so I have argued. This yields a more nuanced view than that suggested by either Laband and Piette (1994), who focus only on connection bias, or by the argument for triple-blinding based on the consequences for the readers, which focuses only on identity bias.

However, in mathematics and some of the humanities there is arguably no positive quality effect from connection bias, as knowing about an author's other work is not taken to be relevant (Easwaran 2009). So here the negative effect of identity bias is the only relevant consideration from the readers' perspective. In this situation, considerations concerning the consequences for the author and considerations concerning the consequences for the readers point in the same direction: in favor of triple-blind review.

A The Acceptance Probability and the Average Quality of Papers

Proposition 2. *$\mu_i^U \sim N(\mu, \sigma_U^2)$ and $\mu_i^K \sim N(\mu, \sigma_K^2)$. Moreover, $\sigma_U^2 < \sigma_K^2$ whenever $\sigma_{sc}^2 > 0$ and $\sigma_{rv}^2 > 0$.*

Proof. First consider the distribution of r_i. Since $r_i \mid q_i \sim N(q_i, \sigma_{rv}^2)$, $q_i \mid \mu_i \sim N(\mu_i, \sigma_{qu}^2)$, and $\mu_i \sim N(\mu, \sigma_{sc}^2)$, it follows that $r_i \mid \mu_i \sim N(\mu_i, \sigma_{qu}^2 + \sigma_{rv}^2)$ and $r_i \sim N(\mu, \sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)$.

The latter can be used straightforwardly to determine the distribution of μ_i^U. Since $r_i - \mu \sim N(0, \sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)$ it follows that

$$\frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}(r_i - \mu) \sim N\left(0, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right) \sim N\left(0, \sigma_U^2\right).$$

The result follows because μ is a constant and

$$\mu_i^U = \frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\mu = \frac{\sigma_{qu}^2 + \sigma_{sc}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}(r_i - \mu) + \mu.$$

Determining the distribution of μ_i^K is slightly trickier because there are two random variables involved: r_i and μ_i. As noted above, $r_i \mid \mu_i \sim N(\mu_i, \sigma_{qu}^2 + \sigma_{rv}^2)$. Thus, writing $X_i = \frac{\sigma_{qu}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}(r_i - \mu_i)$,

$$X_i \mid \mu_i \sim N\left(0, \frac{\sigma_{qu}^4}{\sigma_{qu}^2 + \sigma_{rv}^2}\right).$$

Since

$$\mu_i^K = \frac{\sigma_{qu}^2}{\sigma_{qu}^2 + \sigma_{rv}^2} r_i + \frac{\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\mu_i = X_i + \mu_i$$

it remains to determine the convolution of X_i and μ_i. This can be done using

the moment-generating function and the law of total expectation. Recall that the moment-generating function of an $N(m, s^2)$ distribution is given by $M(t) = \exp\{mt + \frac{1}{2}s^2t^2\}$. So the moment-generating function of μ_i^K is

$$\begin{aligned}
\mathbb{E}[\exp\{t\mu_i^K\}] &= \mathbb{E}[\exp\{t(X_i + \mu_i)\}] \\
&= \mathbb{E}[\mathbb{E}[\exp\{tX_i + t\mu_i\} \mid \mu_i]] \\
&= \mathbb{E}[\exp\{t\mu_i\}\mathbb{E}[\exp\{tX_i\} \mid \mu_i]] \\
&= \exp\left\{0t + \frac{1}{2}\frac{\sigma_{qu}^4}{\sigma_{qu}^2 + \sigma_{rv}^2}t^2\right\}\mathbb{E}[\exp\{t\mu_i\}] \\
&= \exp\left\{\frac{1}{2}\frac{\sigma_{qu}^4}{\sigma_{qu}^2 + \sigma_{rv}^2}t^2 + \mu t + \frac{1}{2}\sigma_{sc}^2t^2\right\} \\
&= \exp\left\{\mu t + \frac{1}{2}\frac{\sigma_{qu}^4 + \sigma_{sc}^2(\sigma_{qu}^2 + \sigma_{rv}^2)}{\sigma_{qu}^2 + \sigma_{rv}^2}t^2\right\},
\end{aligned}$$

which is exactly the moment-generating function of the desired normal distribution.

Finally, note that

$$\begin{aligned}
\sigma_U^2 &= \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2(\sigma_{qu}^2 + \sigma_{rv}^2)}{(\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)(\sigma_{qu}^2 + \sigma_{rv}^2)}, \\
\sigma_K^2 &= \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)^2(\sigma_{qu}^2 + \sigma_{rv}^2) + \sigma_{sc}^2\sigma_{rv}^4}{(\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2)(\sigma_{qu}^2 + \sigma_{rv}^2)}.
\end{aligned}$$

So $\sigma_U^2 < \sigma_K^2$ whenever $\sigma_{sc}^2 > 0$ and $\sigma_{rv}^2 > 0$ (and $\sigma_U^2 = \sigma_K^2$ otherwise, assuming the expressions are well-defined in that case). □

Theorem 3. $\Pr\left(\mu_i^K > q^*\right) > \Pr\left(\mu_i^U > q^*\right)$ *if* $q^* > \mu$, $\sigma_{sc}^2 > 0$, *and* $\sigma_{rv}^2 > 0$.

Proof. It follows from proposition 2 that

$$\Pr\left(\mu_i^K > q^*\right) = 1 - \Phi\left(\frac{q^* - \mu}{\sigma_K}\right) \text{ and } \Pr\left(\mu_i^U > q^*\right) = 1 - \Phi\left(\frac{q^* - \mu}{\sigma_U}\right),$$

where Φ is the distribution function (or cumulative density function) of a standard normal distribution. Since Φ is (strictly) increasing in its argument, and $\sigma_K > \sigma_U$ by proposition 2, the theorem follows immediately. □

In order to prove proposition 4 a number of intermediate results are needed.

Lemma 8.

$$\mathbb{E}[q_i \mid \mu_i^U > q^*] = \mathbb{E}[\mu_i^U \mid \mu_i^U > q^*],$$
$$\mathbb{E}[q_i \mid \mu_i^K > q^*] = \mathbb{E}[\mu_i^K \mid \mu_i^K > q^*].$$

Proof. Because μ_i^U is simply an (invertible) transformation of r_i, it follows that

$$q_i \mid \mu_i^U \sim q_i \mid r_i \sim N\left(\mu_i^U, \frac{(\sigma_{qu}^2 + \sigma_{sc}^2)\sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}\right).$$

The distribution of $q_i \mid \mu_i^K$ is a little trickier to find, because μ_i^K is a linear combination of two random variables, r_i and μ_i, and it is not obvious that learning μ_i^K is as informative as learning both r_i and μ_i. But using the known distributions of $q_i \mid \mu_i$ and $\mu_i^K \mid q_i, \mu_i$ and integrating out μ_i it can be shown that

$$q_i \mid \mu_i^K \sim q_i \mid r_i, \mu_i \sim N\left(\mu_i^K, \frac{\sigma_{qu}^2 \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}\right).$$

The important point here is that $\mathbb{E}[q_i \mid \mu_i^x] = \mu_i^x$ both for $x = U$ and $x = K$.

Now the law of total expectation can be used to establish that

$$\mathbb{E}[q_i \mid \mu_i^x > q^*] = \mathbb{E}[\mathbb{E}[q_i \mid \mu_i^x] \mid \mu_i^x > q^*] = \mathbb{E}[\mu_i^x \mid \mu_i^x > q^*],$$

for $x = U, K$. □

Let $X \sim N(\mu, \sigma^2)$ be a normally distributed random variable. Then $X \mid X > a$ follows a *left-truncated normal distribution*, with left-truncation point a. As a result of lemma 8 I am interested in the mean of left-truncated normal distributions. According to, e.g., Johnson et al. (1994, chapter 13, section 10.1), this mean can be expressed as

$$\mathbb{E}[X \mid X > a] = \mu + \sigma R\left(\frac{a-\mu}{\sigma}\right). \tag{1}$$

Here

$$R(x) = \frac{\phi(x)}{1 - \Phi(x)}$$

for all $x \in \mathbb{R}$, where ϕ is the probability density function of the standard normal distribution, and Φ is its distribution function. R is the inverse of what is known in the literature (e.g., Gordon 1941) as *Mills' ratio*.

It follows from the definitions that $R(x) > 0$ for all $x \in \mathbb{R}$ and that

$$R'(x) = R(x)^2 - xR(x). \tag{2}$$

Proposition 9 (Gordon (1941))**.** *For all $x > 0$, $R(x) < \frac{x^2+1}{x}$.*

Proposition 9 can be used to establish the next result.

Proposition 10. *If $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\mu, s^2)$ with $s > \sigma > 0$ then $\mathbb{E}[Y \mid Y > a] > \mathbb{E}[X \mid X > a]$.*

Proof. It suffices to show that the derivative $\frac{\partial}{\partial \sigma}\mathbb{E}[X \mid X > a]$ is positive for all $\sigma > 0$. Differentiating equation (1) (using equation (2)) yields

$$\frac{\partial}{\partial\sigma}\mathbb{E}[X \mid X > a] = \left(\left(\frac{a-\mu}{\sigma}\right)^2 + 1\right) R\left(\frac{a-\mu}{\sigma}\right) - \frac{a-\mu}{\sigma} R\left(\frac{a-\mu}{\sigma}\right)^2.$$

Since $R\left(\frac{a-\mu}{\sigma}\right) > 0$, $\frac{\partial}{\partial\sigma}\mathbb{E}[X \mid X > a] > 0$ if and only if

$$\left(\frac{a-\mu}{\sigma}\right)^2 + 1 - \frac{a-\mu}{\sigma} R\left(\frac{a-\mu}{\sigma}\right) > 0.$$

This is true whenever $\frac{a-\mu}{\sigma} \leq 0$ because then both terms in the sum are positive. Proposition 9 guarantees that it is true whenever $\frac{a-\mu}{\sigma} > 0$ as well. □

Proposition 4. $\mathbb{E}[q_i \mid \mu_i^K > q^*] > \mathbb{E}[q_i \mid \mu_i^U > q^*]$ *whenever* $\sigma_{sc}^2 > 0$, *and* $\sigma_{rv}^2 > 0$.

Proof. By lemma 8,

$$\mathbb{E}[q_i \mid \mu_i^U > q^*] = \mathbb{E}[\mu_i^U \mid \mu_i^U > q^*],$$
$$\mathbb{E}[q_i \mid \mu_i^K > q^*] = \mathbb{E}[\mu_i^K \mid \mu_i^K > q^*].$$

By proposition 2, $\mu_i^U \sim N(\mu, \sigma_U^2)$ and $\mu_i^K \sim N(\mu, \sigma_K^2)$, with $\sigma_U < \sigma_K$. Hence the conditions of proposition 10 are satisfied, and the result follows. □

Proposition 6.

$$\mu_i^{KA} \sim N\left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2\right),$$
$$\mu_i^{KF} \sim N\left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \sigma_K^2\right),$$
$$\mu_i^{UA} \sim N\left(\mu - \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2\right),$$
$$\mu_i^{UF} \sim N\left(\mu + \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \sigma_U^2\right).$$

Proof. Since μ_i^{KA} and μ_i^{KF} are simply μ_i^K shifted by a constant (see proposition 5) they follow the same distribution as μ_i^K except that its mean is shifted by the same constant. Similarly μ_i^{UA} and μ_i^{UF} are just μ_i^U shifted by a constant. So the results follow from proposition 2. □

For notational convenience, I introduce q^{KA}, q^{KF}, q^{UA}, and q^{UF}, defined by

$$q^{KA} = q^* + \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2}, \qquad q^{KF} = q^* - \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{rv}^2},$$
$$q^{UA} = q^* + \frac{\varepsilon \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}, \qquad q^{UF} = q^* - \frac{\delta \cdot \sigma_{rv}^2}{\sigma_{qu}^2 + \sigma_{sc}^2 + \sigma_{rv}^2}.$$

Theorem 7. *If $\varepsilon > 0$, $\delta > 0$, and $\sigma_{rv}^2 > 0$,*

$$\Pr\left(\mu_i^{KA} > q^*\right) < \Pr\left(\mu_i^{KF} > q^*\right) \text{ and } \Pr\left(\mu_i^{UA} > q^*\right) < \Pr\left(\mu_i^{UF} > q^*\right).$$

Proof. For the first inequality, note that

$$\Pr\left(\mu_i^{KA} > q^*\right) = 1 - \Phi\left(\frac{q^{KA} - \mu}{\sigma_K}\right) < 1 - \Phi\left(\frac{q^{KF} - \mu}{\sigma_K}\right) = \Pr\left(\mu_i^{KF} > q^*\right).$$

The equalities follow from the distributions of the posterior means established in proposition 6. The inequality follows from the fact that Φ is strictly increasing in its argument. By the same reasoning,

$$\Pr\left(\mu_i^{UA} > q^*\right) = 1 - \Phi\left(\frac{q^{UA} - \mu}{\sigma_U}\right) < 1 - \Phi\left(\frac{q^{UF} - \mu}{\sigma_U}\right) = \Pr\left(\mu_i^{UF} > q^*\right).$$

□

Lemma 11.

$$\begin{aligned}\Pr(A_i) &= p_{KA}\left(1-\Phi\left(\frac{q^{KA}-\mu}{\sigma_K}\right)\right)+p_{KF}\left(1-\Phi\left(\frac{q^{KF}-\mu}{\sigma_K}\right)\right)\\ &\quad+p_{UA}\left(1-\Phi\left(\frac{q^{UA}-\mu}{\sigma_U}\right)\right)+p_{UF}\left(1-\Phi\left(\frac{q^{UF}-\mu}{\sigma_U}\right)\right).\\ \mathbb{E}[q_i \mid A_i] &= \mu+\frac{\sigma_K}{\Pr(A_i)}\left(p_{KA}\phi\left(\frac{q^{KA}-\mu}{\sigma_K}\right)+p_{KF}\phi\left(\frac{q^{KF}-\mu}{\sigma_K}\right)\right)\\ &\quad+\frac{\sigma_U}{\Pr(A_i)}\left(p_{UA}\phi\left(\frac{q^{UA}-\mu}{\sigma_U}\right)+p_{UF}\phi\left(\frac{q^{UF}-\mu}{\sigma_U}\right)\right).\end{aligned}$$

Proof. The expression for $\Pr(A_i)$ follows immediately from the distributions of the posterior means established in proposition 6.

To get an expression for $\mathbb{E}[q_i \mid A_i]$, consider first the average quality of scientist i's paper given that it is accepted and given that scientist i is in the group of scientists known to the editor that the editor is biased against. This average quality is

$$\begin{aligned}\mathbb{E}\left[q_i \mid \mu_i^{KA} > q^*\right] &= \mathbb{E}\left[q_i \mid \mu_i^{K} > q^{KA}\right] = \mathbb{E}\left[\mu_i^{K} \mid \mu_i^{K} > q^{KA}\right]\\ &= \mu+\sigma_K R\left(\frac{q^{KA}-\mu}{\sigma_K}\right),\end{aligned}$$

where the first equality simply rewrites the inequality $\mu_i^{KA} > q^*$ in a more convenient form, the second equality uses lemma 8, and the third equality uses equation 1. Similarly,

$$\mathbb{E}\left[q_i \mid \mu_i^{KF} > q^*\right] = \mu + \sigma_K R\left(\frac{q^{KF}-\mu}{\sigma_K}\right),$$
$$\mathbb{E}\left[q_i \mid \mu_i^{UA} > q^*\right] = \mu + \sigma_U R\left(\frac{q^{UA}-\mu}{\sigma_U}\right),$$
$$\mathbb{E}\left[q_i \mid \mu_i^{UF} > q^*\right] = \mu + \sigma_U R\left(\frac{q^{UF}-\mu}{\sigma_U}\right).$$

The average quality of accepted papers $\mathbb{E}[q_i \mid A_i]$ is a weighted sum of these expectations. The weights are given by the proportion of accepted papers that are written by a scientist in that particular group. For example, authors known to the editor that she is biased against form a $p_{KA}\Pr(\mu_i^{KA} > q^*)/\Pr(A_i)$ proportion of accepted papers. Hence

$$\begin{aligned}
\mathbb{E}\left[q_i \mid A_i\right] = {} & \frac{1}{\Pr\left(A_i\right)} p_{KA} \Pr\left(\mu_i^{KA} > q^*\right) \mathbb{E}\left[q_i \mid \mu_i^{KA} > q^*\right] \\
& + \frac{1}{\Pr\left(A_i\right)} p_{KF} \Pr\left(\mu_i^{KF} > q^*\right) \mathbb{E}\left[q_i \mid \mu_i^{KF} > q^*\right] \\
& + \frac{1}{\Pr\left(A_i\right)} p_{UA} \Pr\left(\mu_i^{UA} > q^*\right) \mathbb{E}\left[q_i \mid \mu_i^{UA} > q^*\right] \\
& + \frac{1}{\Pr\left(A_i\right)} p_{UF} \Pr\left(\mu_i^{UF} > q^*\right) \mathbb{E}\left[q_i \mid \mu_i^{UF} > q^*\right] \\
= {} & \mu + \frac{\sigma_K}{\Pr\left(A_i\right)} \left(p_{KA}\phi\left(\frac{q^{KA}-\mu}{\sigma_K}\right) + p_{KF}\phi\left(\frac{q^{KF}-\mu}{\sigma_K}\right)\right) \\
& + \frac{\sigma_U}{\Pr\left(A_i\right)} \left(p_{UA}\phi\left(\frac{q^{UA}-\mu}{\sigma_U}\right) + p_{UF}\phi\left(\frac{q^{UF}-\mu}{\sigma_U}\right)\right). \qquad \square
\end{aligned}$$

References

Charles D. Bailey, Dana R. Hermanson, and Timothy J. Louwers. An examination of the peer review process in accounting journals. *Jour-*

nal of Accounting Education, 26(2):55–72, 2008a. ISSN 0748-5751. doi: 10.1016/j.jaccedu.2008.04.001. URL http://www.sciencedirect.com/science/article/pii/S0748575108000201.

Charles D. Bailey, Dana R. Hermanson, and James G. Tompkins. The peer review process in finance journals. *Journal of Financial Education*, 34: 1–27, 2008b. ISSN 0093-3961. URL http://www.jstor.org/stable/41948838.

Damien Besancenot, Kim V. Huynh, and Joao R. Faria. Search and research: the influence of editorial boards on journals' quality. *Theory and Decision*, 73(4):687–702, 2012. ISSN 0040-5833. doi: 10.1007/s11238-012-9314-7. URL http://dx.doi.org/10.1007/s11238-012-9314-7.

Liam Kofi Bright. Against candidate quality. Manuscript, 2015. URL https://www.academia.edu/11673059/Against_Candidate_Quality.

Liam Kofi Bright, Daniel Malinsky, and Morgan Thompson. Causally interpreting intersectionality theory. *Philosophy of Science*, 83(1):60–81, 2016. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/684173.

Justin Bruner and Cailin O'Connor. Power, bargaining, and collaboration. In Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg, editors, *Scientific Collaboration and Collective Knowledge*. Oxford University Press, Oxford, forthcoming. URL http://philpapers.org/rec/BRUPBA-2.

Amber E. Budden, Tom Tregenza, Lonnie W. Aarssen, Julia Koricheva, Roosa Leimu, and Christopher J. Lortie. Double-blind review favours increased representation of female authors. *Trends in Ecology & Evolution*, 23(1):4–6, 2008. ISSN 0169-5347. doi: 10.1016/j.tree.2007.07.008. URL http://www.sciencedirect.com/science/article/pii/S0169534707002704.

Kenneth E. Clark. *America's Psychologists: A Survey of a Growing Profession*. American Psychological Association, Washington, 1957.

Jonathan R. Cole and Stephen Cole. Measuring the quality of sociological research: Problems in the use of the "Science Citation Index". *The American Sociologist*, 6(1):23–29, 1971. ISSN 00031232. URL http://www.jstor.org/stable/27701705.

Stephen Cole and Jonathan R. Cole. Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review*, 32(3):377–390, 1967. ISSN 00031224. URL http://www.jstor.org/stable/2091085.

Stephen Cole and Jonathan R. Cole. Visibility and the structural bases of awareness of scientific research. *American Sociological Review*, 33(3): 397–413, 1968. ISSN 00031224. URL http://www.jstor.org/stable/2091914.

Patricia Hill Collins and Valerie Chepp. Intersectionality. In Georgina Waylen, Karen Celis, Johanna Kantola, and S. Laurel Weldon, editors, *The Oxford Handbook of Gender and Politics*, chapter 2, pages 57–87. Oxford University Press, Oxford, 2013. ISBN 0199751455.

Rick Crandall. Editorial responsibilities in manuscript review. *Behavioral and Brain Sciences*, 5:207–208, Jun 1982. ISSN 1469-1825. doi: 10.1017/S0140525X00011316. URL http://journals.cambridge.org/article_S0140525X00011316.

Diana Crane. The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist*, 2(4):195–201, 1967. ISSN 00031232. URL http://www.jstor.org/stable/27701277.

Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, New Jersey, 2004.

Kenny Easwaran. Probabilistic proofs and transferability. *Philosophia Mathematica*, 17(3):341–362, 2009. doi: 10.1093/philmat/nkn032. URL http://philmat.oxfordjournals.org/content/17/3/341.abstract.

Glenn Ellison. Evolving standards for academic publishing: A q-r theory. *Journal of Political Economy*, 110(5):994–1034, 2002. ISSN 00223808. URL http://www.jstor.org/stable/10.1086/341871.

João Ricardo Faria. The game academics play: Editors versus authors. *Bulletin of Economic Research*, 57(1):1–12, 2005. ISSN 1467-8586. doi: 10.1111/j.1467-8586.2005.00212.x. URL http://dx.doi.org/10.1111/j.1467-8586.2005.00212.x.

Miranda Fricker. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, Oxford, 2007.

Robert D. Gordon. Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941. ISSN 00034851. URL http://www.jstor.org/stable/2235868.

Remco Heesen. Academic superstars: Competent or lucky? *Synthese*, forthcoming. doi: 10.1007/s11229-016-1146-5. URL http://dx.doi.org/10.1007/s11229-016-1146-5.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. John Wiley & Sons, New York, second edition, 1994.

Philip Kitcher. *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford University Press, Oxford, 1993. ISBN 0195046285.

David N. Laband. Publishing favoritism: A critique of department rankings based on quantitative publishing performance. *Southern Economic Journal*, 52(2):510–515, 1985. ISSN 00384038. URL http://www.jstor.org/stable/1059636.

David N. Laband and Michael J. Piette. Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors. *Journal of Political Economy*, 102(1):194–203, 1994. ISSN 00223808. URL http://www.jstor.org/stable/2138799.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17, 2013. ISSN 1532-2890. doi: 10.1002/asi.22784. URL http://dx.doi.org/10.1002/asi.22784.

D. Lindsey. Using citation counts as a measure of quality in science: Measuring what's measurable rather than what's valid. *Scientometrics*, 15 (3–4):189–203, 1989. ISSN 0138-9130. doi: 10.1007/BF02017198. URL http://dx.doi.org/10.1007/BF02017198.

Christopher D. Mackie. *Canonizing Economic Theory: How Theories and Ideas Are Selected in Economics*. M. E. Sharpe, New York, 1998. ISBN 9780765602848.

Conor Mayo-Wilson, Kevin J. S. Zollman, and David Danks. The independence thesis: When individual and social epistemology diverge. *Philosophy of Science*, 78(4):653–677, 2011. ISSN 00318248. URL http://www.jstor.org/stable/10.1086/661777.

Marshall H. Medoff. Editorial favoritism in economics? *Southern Economic Journal*, 70(2):425–434, 2003. ISSN 00384038. URL http://www.jstor.org/stable/3648979.

Robert K. Merton. A note on science and democracy. *Journal of Legal and Political Sociology*, 1(1–2):115–126, 1942. Reprinted in Merton (1973, chapter 13).

Robert K. Merton. *The Sociology of Science: Theoretical and Empirical Investigations*. The University of Chicago Press, Chicago, 1973. ISBN 0226520919.

Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41): 16474–16479, 2012. doi: 10.1073/pnas.1211286109. URL `http://www.pnas.org/content/109/41/16474.abstract`.

Michael J. Piette and Kevin L. Ross. A study of the publication of scholarly output in economics journals. *Eastern Economic Journal*, 18(4):429–436, 1992. ISSN 00945056. URL `http://www.jstor.org/stable/40325474`.

Jennifer Saul. Implicit bias, stereotype threat, and women in philosophy. In Katrina Hutchison and Fiona Jenkins, editors, *Women in Philosophy: What Needs to Change?*, chapter 2, pages 39–60. Oxford University Press, Oxford, 2013.

Daniel L. Sherrell, Joseph F. Hair, Jr., and Mitch Griffin. Marketing academicians' perceptions of ethical research and publishing behavior. *Journal of the Academy of Marketing Science*, 17(4):315–324, 1989. ISSN 0092-0703. doi: 10.1007/BF02726642. URL `http://dx.doi.org/10.1007/BF02726642`.

Kenneth J. Smith and Robert F. Dombrowski. An examination of the relationship between author-editor connections and subsequent citations of auditing research articles. *Journal of Accounting Education*, 16 (3–4):497–506, 1998. ISSN 0748-5751. doi: 10.1016/S0748-5751(98)

00019-0. URL http://www.sciencedirect.com/science/article/pii/S0748575198000190.

Rhea E. Steinpreis, Katie A. Anders, and Dawn Ritzke. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, 41(7–8):509–528, 1999. ISSN 0360-0025. doi: 10.1023/A:1018839203698. URL http://dx.doi.org/10.1023/A:1018839203698.

Brandon D. Stewart and B. Keith Payne. Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, 34(10):1332–1345, 2008. doi: 10.1177/0146167208321269. URL http://psp.sagepub.com/content/34/10/1332.abstract.

Michael Strevens. The role of the priority rule in science. *The Journal of Philosophy*, 100(2):55–79, 2003. ISSN 0022362X. URL http://www.jstor.org/stable/3655792.

Eric Luis Uhlmann and Geoffrey L. Cohen. "I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, 104(2):207–223, 2007. ISSN 0749-5978. doi: 10.1016/j.obhdp.2007.07.001. URL http://www.sciencedirect.com/science/article/pii/S0749597807000611.

Virginia Valian. *Why So Slow? The Advancement of Women*. MIT Press, Cambridge, 1999. ISBN 9780262720311.

Christine Wennerås and Agnes Wold. Nepotism and sexism in peer-review. *Nature*, 387(6631):341–343, May 1997. ISSN 0028-0836. doi: 10.1038/387341a0. URL http://dx.doi.org/10.1038/387341a0.

Alan J. Ziobrowski and Karen M. Gibler. Factors academic real estate authors consider when choosing where to submit a manuscript for pub-

lication. *Journal of Real Estate Practice and Education*, 3(1):43–54, 2000. ISSN 1521-4842. URL http://ares.metapress.com/content/1762151051KM2227.

Kevin J. S. Zollman. Optimal publishing strategies. *Episteme*, 6:185–199, Jun 2009. ISSN 1750-0117. doi: 10.3366/E174236000900063X. URL http://journals.cambridge.org/article_S1742360000001283.

Atlanta, GA; 3-5 November 2016 -255-

Strategies of Explanatory Abstraction in Molecular Systems Biology[†]

Nicholaos Jones[‡]

Abstract

I consider three explanatory strategies from recent systems biology that are driven by mathematics as much as mechanistic detail. Analysis of differential equations drives the first strategy; topological analysis of network motifs drives the second; mathematical theorems from control engineering drive the third. I also distinguish three abstraction types: aggregations, which simplify by condensing information; generalizations, which simplify by generalizing information; and structurations, which simplify by contextualizing information. Using a common explanandum as reference point—namely, the robust perfect adaptation of chemotaxis in *Escherichia coli*—I argue that each strategy invokes a different combination of abstraction types and that each targets its abstractions to different mechanistic details.

1 Introductory Remarks

The currently dominant paradigm for understanding explanation in biology puts mechanism at center stage (Nicholson 2012; Levy 2013). Leading accounts of mechanistic explanation, while differing in the particulars of their analysis of *mechanism*, agree that mechanistic explanations explain by alluding to mechanisms or models thereof (Machamer, Darden, Craver 2000; Bechtel and Abrahamsen 2005).

There is a small publishing industry devoted to discerning the scope of mechanistic explanation in scientific practice. Some claim to identify biological explanations that do not allude to mechanisms (Wouters 2007; Huneman 2010; Rice 2015). Fans of mechanistic explanation tend to resist making scope concessions, preferring instead to accommodate the putative explanations as mechanistic despite initial appearances, to broaden the scope of mechanistic explanation or the analysis of *mechanism*, or else to

[†] Draft. For symposium on *Integrating Explanatory Strategies Across the Life Sciences* at the 2016 meeting of Philosophy of Science Association, Atlanta, GA. I thank audiences at Mississippi State University, the Alabama Philosophical Society, and the Society for Philosophy of Science in Practice for comments on earlier drafts.

[‡] Department of Philosophy, University of Alabama in Huntsville, Huntsville AL 35899, nick.jones@uah.edu

deny that the putative explanations are explanations at all (Craver 2006; Bechtel and Abrahamsen 2010; Brigandt 2013; Levy and Bechtel 2013).

I set aside questions about what qualifies as an explanation as well as questions about whether only mechanisms—or models thereof—carry explanatory power. I focus, instead, on *explanatory strategies*, understood as patterns of reasoning directed toward providing explanations. I consider three explanatory strategies from recent systems biology that are driven by mathematics as much as, if not more than, mechanistic detail. Analysis of differential equations drives the first strategy; topological analysis of network motifs drives the second; mathematical theorems from control engineering drive the third.

Systems biologists use these strategies to supplement the explanatory power of traditional molecular mechanisms (see Brigandt et al *forthcoming*). My aim is to identify how the strategies differ from each other, rather than how they differ from standard mechanistic explanations or what might unify them in those differences (for which see Green and Jones 2016). Doing so helps with understanding relations among the strategies, their tactics for integrating mechanistic detail, and explanatory affordances of their mathematical elements.

They key to my analysis is a distinction among three abstraction types: aggregations, which simplify by condensing information; generalizations, which simplify by generalizing information; and structurations, which simplify by contextualizing information. Using a common explanandum as reference point—namely, the robust perfect adaptation of chemotaxis in *Escherichia coli* (Barkai and Leibler 1997; Ma et al 2009; Yi et al 2000)—I argue that each strategy invokes a different combination of abstraction types and that each targets its abstractions to different mechanistic details. I begin with the typology of abstraction.

2 Abstraction Typology

I am interested in abstractions as representational rather than metaphysical. Abstractions, as I understand them, are ontologically innocent, so that characterizing features of representations as abstractions over some parts of reality carries no implication that features correspond to abstract objects (see also Cartwright 1989, 353-354; Levy and Bechtel 2013, 243). So, for example, representing the relation between a person, a hotel, and a date range as a reservation does not entail that some abstract object, a *reservation*, exists; nor does representing the motions of an object's constituents as the motion of the object's center of mass entail that some abstract object, a *center of mass*, exists.

Levy and Bechtel characterize a representation as abstract insofar as a more concrete representation is possible (2013, 242). Brigandt and colleagues suggest that biologists use abstractions to "elucidate system-level patterns of organization that may not be visible at the level of molecular details" (*forthcoming*). I concur. I understand abstractions as representing only some of the many elements—objects, relations, parameters—associated with their targets, thereby making apparent patterns obscured by more detailed representations. I add to these insights that biologists produce (at least) three types of abstraction.

Following Ordorica, I call the first *aggregation* (2015, 163-164). An aggregation represents some relationship among multiple elements of a representational target as a higher-level object, or multiple elements of the target as a single, composite object. (See Figure 1a.) Paradigm cases of aggregations include representations of person-hotel-date relations as *reservations*; of costs of services and costs of goods as *costs*; and of the motions of an object's parts as the *motion of a center of mass* (from Ordorica 2015, 164). Aggregations abstract from plurality to individual, ignoring differences among many in order to make salient some integrated unity among the elements of a representational target. They thereby simplify representations by condensing information about representational targets.

Following Pincock, I call the second abstraction type *generalization* (2015, 864). A generalization represents some element of a representational target as a class of elements, where potential instances of the class might include elements not present in the target. (See Figure 1b.) For example, because the class of solution measures includes all soap-bubble-like surfaces, such as the cellular froth surrounding radiolarian protozoa, representing a soap-bubble surface as a "solution measure" is a generalization (Pincock 2015, 864). Generalizations abstract from an instance to a class thereof, ignoring differences between instances of the class in order to make salient some more general unity. They thereby simplify representations by generalizing from information about representational targets.

I call the third abstraction type *structuration*. A structuration represents some element of a representational target as a position in a structure, such that potential occupants of the position might include elements not present in the target. (See Figure 1c.) I follow Haslanger in understanding structures as "complex entities with parts whose behavior is constrained by their relation to other parts" (2016, 118). Paradigm cases of structurations include representating Barack Obama as President of the United States of America, or representing AIneias as son of Anchises and Aphrodite. Structurations abstract to a position in a structure, from an occupant of the position, ignoring intrinsic features of the occupant unrelated to its position in order to make salient the

occupant's role relative to occupants of other positions in the same structure. They thereby simplify by contextualizing information about representational targets.

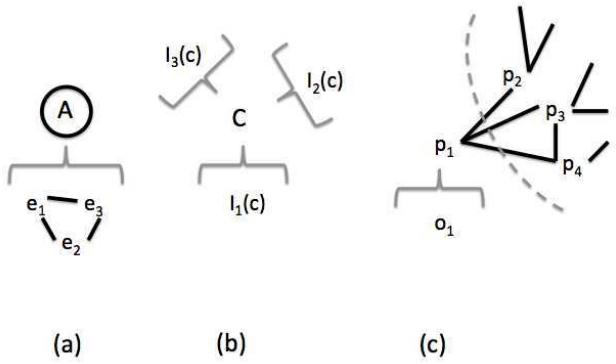


Figure 1: Visualizing Abstraction Types. (a) Aggregation A represents elements e1, e2, and e3 (and relations therein) as a single object. (b) Generalization C represents $I_1(c)$ as a class, instances of which also include $I_2(c)$ and $I_3(c)$. (c) Structuration p_1 represents element o_1 as a position in larger structure that also includes p_2, p_3, and p_4.

I understand aggregations as distinct from both generalizations and structurations, by virtue of being many-to-one, rather than one-to-one, simplifications. I also understand being a generalization as insufficient for being a structuration. For representations of positions carry information about functional relationships between their occupants and other positions in the same structure; but representations of classes do not. Finally, insofar as classes are sets, I understand being a structuration as insufficient for being a generalization. For, sometimes, representing target elements as classes carries some information about intrinsic features of those elements apart from their functional relations to elements occupying other positions in the same structure; but representing target elements as positions in structures never carries such information.

3 Robust Perfect Adaptation of *E.coli* Chemotaxis

My central claim is that different explanatory strategies from recent systems biology differ from each other, at least in part, by virtue of appealing to different abstraction types. I support this claim by considering a case in which multiple strategies target the same explanandum. Doing so minimizes confounds that confuse differences due to the nature of each explanatory strategy with differences due to the nature of each

explanatory target. I focus on a particular explanandum known as robust perfect adaptation of bacterial chemotaxis, following others who consider this a paradigmatic target for non-mechanistic explanation (Braillard 2010; Brigandt, Green, and O'Malley *forthcoming*; Matthiessen *forthcoming*).

3.1 Explanadum Context

Escherichia coli (*E.coli*) is popular model organism in biological research. It is very sensitive to small chemical changes over a very large range of background concentrations. It also has a simple and well-understood signal transduction network (Wadhams and Armitage 2004).

E.coli manages two kinds of motion (Berg 2003). It *runs* by rotating its flagellar motor counterclockwise. This aligns all of its flagella into a synchronized bundle, resulting in movement in a straight line for about 1 second. *E.coli* also *tumbles* by rotating its flagellar motor clockwise. This breaks flagellar alignment, and the asynchronized flagella produce stationary changes of direction lasting for about 0.1 second. *E.coli* are randomly reoriented after each tumble. Moreover, while these tumbles occur with regular frequency, *E.coli* with higher concentrations of CheR protein tumble more frequently (Spudich and Kochland 1975).

E.coli's motion in a uniform external environment resembles a random walk. *E.coli* has no ability to control or select its direction of motion, and its straight runs are subject to Brownian motion because of eddies. However, in the presence of a chemical attractant—amino acids such as serine or aspartic acid, or sugars such as maltose or glucose— *E.coli taxis* toward the attractant. This taxi behavior involves less frequent tumbles, leading to longer runs and so gradual motion toward the attractant. (There is an opposite behavior for repellants such as metal ions or leucine.)

The biomolecular mechanism for *E.coli* chemotaxis is well-understood. When an environmental attractant attaches to a receptor, the receptor lowers the activity of the CheW-CheA protein complex. Less activity from this complex reduces the rate of CheY phosphorylation, which results in less phorphorylated CheY diffusing to the flagella. Because CheY induces clockwise rotation of the flagellar motor, the outcome is less frequent tumbling.

3.2 Explanadum Question

Alon and colleagues have experimental verification that, in the presence of a chemical attractant mixed uniformly into the environment at a constant concentration, *E.coli* chemotaxis *perfectly adaptive* (Alon et al 2009). After a brief period of decreased tumbling frequency, the frequency of *E.coli* tumbles increases toward and returns to the

exact frequency prior to the introduction of the attractant. The effect of the attractant, accordingly, becomes entirely forgotten despite its continuing presence.

The biomolecular mechanism for the adaptiveness of chemotaxis for *E.coli* is also well-understood. Some time after a new attractant has been detected by receptors, the lower activity of the CheW-CheA complex induces less CheB activity. This reduces the rate for removing methyl groups from the CheW-CheA complex and, together with continual methylation of the CheR receptor, CheW-CheA methylation increases. More methylation means more CheW-CheA activity, which in turn induces more CheY phosphorylation. This eventually results in more phosphorylated CheY diffusing to the flagellar motor, which increases clockwise motor rotation and thereby raises tumbling frequency.

Alon and colleagues have further experimental verification that this perfectly adaptive chemotaxis of *E.coli* is *robust* across ranges of CheR concentrations 0.5 to 50 times higher than concentration levels in "wild type" *E.coli* (Alon et al 2009). (By contrast, *E.coli*'s adaptation time—the time to return to 50% of its pre-stimulus tumbling frequency—is not robust to different CheR concentrations, because more CheR entails longer adaptation times.) This is the explanandum of interest: why is the perfect adaptation of *E.coli* chemotaxis, in the presence of a well-distributed chemical attractant, robust to CheR protein concentrations?

There are (at least) three strategies for answering this question in recent systems biology literature. (For a fourth, see Kollman et al 2005.) I consider each in turn, first sketching the general strategy and then making explicit the abstractions at work.

4 Distinguishing Explanatory Strategies through Abstraction Types

4.1 Dynamical Modeling

I call the first strategy *dynamical modeling*. This strategy begins by constructing a chemotaxis network for *E.coli*. This network represents the mechanism for *E.coli* chemotaxis, including specific biochemical details about when and how relevant proteins affect each other. (See Figurer 2.) For example, Barkai and Leibler (1997) construct a model according to which, among many other specifics, CheB demethylates only the active form of the CheW-CheA complex and CheR works only at saturation.

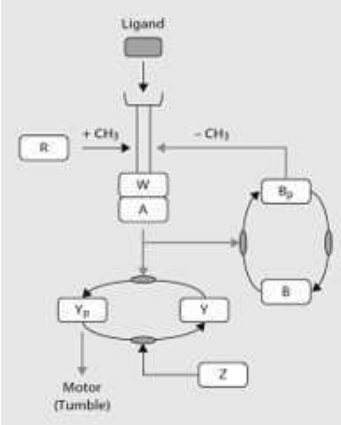


Figure 2. Mechanistic network for *E.coli* chemotaxis (Rao and Ordal 2009).

The dynamical modeling strategy proceeds by constructing a dynamical model—typically a set of differential equations—from the network (see Jones and Wolkenhauer 2012). One then demonstrates, via mathematical proof or simulation, that this model predicts perfect adaptation in the presence of a well-distributed chemical attractant for CheR concentration values varying over several orders of magnitude. (Raerinne 2013 calls this *sensitivity analysis*.) The demonstration supports the inference that *E.coli* chemotaxis exhibits robust perfect adaptation *because of its biochemical specifics*.

Bechtel and Abrahamsen (2010) call the product of this strategy a *dynamical mechanistic explanation*. I set aside the issue of whether the dynamical modeling strategy produces explanations. But I endorse Bechtel and Abrahamsen's insight that the dynamical modeling strategy produces accounts that are mechanistic, by virtue of depending upon mechanistic details, as well as dynamical, by virtue of analyzing mathematical models built upon those details. For example, Barkai and Leibler's (1997) mathematical analysis is relevant to *E.coli* chemotaxis only insofar as their network details are relevant; and analysis of the network apart from the model cannot produce an inference about the *robustness* of *E.coli*'s perfectly adaptive chemotaxis.

Let's treat the dynamical model driving this explanatory strategy as an initial baseline for evaluating the number and severity of abstraction in various explanatory strategies. The model is abstract in various ways. But we shall treat it as a recipient of further abstractions, in the way a vehicle receives freight. Just as we can determine the weight of the freight indirectly by subtracting the gross weight of vehicle and freight from the "tare weight" (the weight of vehicle alone), we shall determine abstraction variety and

severity/extent for models driving other explanatory strategies by "subtracting" their total abstraction variety and severity from the "tare" abstraction.

4.2 Topological Analysis

I call the second explanatory strategy *topological analysis*. This strategy begins by identifying all possible minimal adaptation networks capable of predicting robust perfect adaptation for *E.coli* chemotaxis. These networks, like the networks for dynamical modeling, represent mechanisms for *E.coli* chemotaxis. Yet, unlike the networks for dynamical modeling, these networks are minimal: they contain the fewest possible nodes and links that suffice for robustly perfectly adaptive chemotaxis. The procedure for identifying all possible minimal networks of this sort is brute computational search. It turns out that there are exactly three, each of which has exactly three nodes and no more than three links (Ma et al 2009).

The topological analysis strategy proceeds by identifying a chemotaxis network known to predict robust perfect adaptation. This strategy thereby relies upon the dynamical modeling strategy, but only for mathematical results. The biochemical details of the chosen chemotaxis network turn out to be largely irrelevant, because the topological analysis strategy proceeds by demonstrating that a *reduced form* of the chosen network is topological equivalent to one of the minimal adaptation models. Reduced forms for mechanistic networks functional equivalents for node groups, group nodes or equivalents into modules, and ignore links within modules in favor of links between modules.

Consider, for example, one of the three minimal adaptation networks Ma and colleagues (2009) discover for *E.coli* chemotaxis. (See Figure 3.) The network has an input activating node A, A inhibiting being activated by B, A also activating C, and C activating some output. Ma and colleagues show that Barkai and Leibler's (1997) model for *E.coli* chemotaxis reduces to this minimal network. Barkai and Leibler have an input and CheR activating, and CheB inhibiting, receptors; these receptors activating the CheW-CheA complex; the complex activating CheB and CheY; and CheY activating some output. Ma and colleagues reconceptualize Barkai and Leibler's network into one where the input activates a *receptor complex*; this complex activates CheY, which activates the output; the complex also activates CheB, which inhibits a *methylation level* also activated by CheR.; and this methylation level activates the receptor complex. Then, in a second reconceptualization that produces one of their minimal adaptation networks, they group the receptor complex and CheB into module A, group CheR and the methylation level into module B, and rename CheY module C.

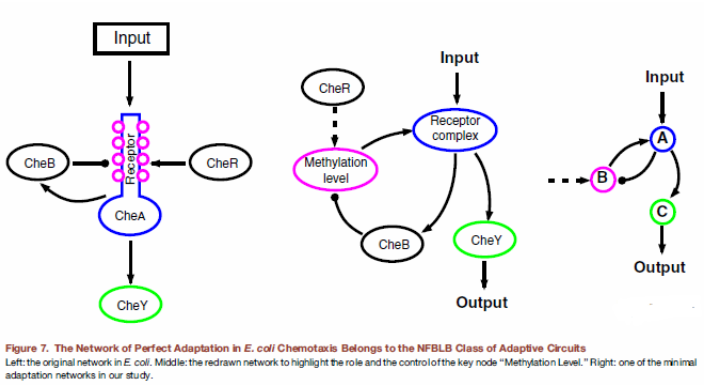


Figure 3: Network topology for *E.coli* chemotaxis (Ma et al 2009).

The topological analysis strategy infers, from the topological equivalence between a minimal adaptation network and the reduced form of a network known to predict robust perfect adaptation for chemotaxis, that *E.coli* chemotaxis exhibits robust perfect adaptation *because of the topology of its chemotaxis network*. Huneman (2010) calls the product of this strategy a *topological explanation*. Regardless of whether analyses such as Ma and colleagues's are explanatory, they are topological by virtue of demonstrating some consequence about the topological properties of a network. This means that, even if the mechanistic details of *E.coli*'s chemotaxis network were different, and even if the biochemical specifics of the network chosen for reduction were different, the product of the topological analysis strategy would remain the same provided that the alternative networks preserve topological equivalence with the originals (see also Jones 2014).

The topological model driving this second explanatory strategy is more abstract than the dynamical model driving our initial ("tare") strategy. The topological model contains more aggregations. For example, it represents CheY and CheZ as "the motor rotation group;" it represents CheA and CheW as "the receptor complex;" and it represents the receptor complex and CheB as "the phosphorylation group." The topological model also contains more structurations. For example, it represents the phosphorylation group as "A" and the motor rotation group as "C." These representations abstract entirely from any intrinsic marks that might distinguish instances of "A" from instances of "C," relying instead upon extrinsic relations to distinguish the nodes from each other. So, for example, "A" but not "C" inhibits "B," "A" activates "C," and so on.

4.3 Organizational Design

I call the third explanatory strategy *organization design*. This strategy begins with a proof to the effect that systems exhibit robust perfect adaptation if and only if they

satisfy the characteristic equation for Integral Feedback Control (IFC). The proof is purely mathematical, well-known from control engineering theory in contexts involving mechanical systems that exhibit IFC such as thermostats. I am not aware of a complete and published version of this proof, but Yi and colleagues (2000) provide a sketch with relevant details. The organizational design strategy proceeds by inferring that *E.coli* chemotaxis exhibits robust perfect adaptation if and only if it satisfies the characteristic equation for IFC, and further inferring that *E.coli* chemotaxis exhibits robust perfect adaptation *because it satisfies the characteristic equation for IFC*. (For better explanatory details regarding this specific case, Braillard 2010; Green and Jones 2016.)

The organizational design strategy invokes neither mechanistic specifics about the chemotaxis network for *E.coli* nor topological details about the structure of that network. The strategy takes the explanandum phenomenon as given, using a mathematical equivalence result to identify a principle both necessary and sufficient for the phenomenon. The strategy thereby has affinities with explanatory strategies that appeal to organizing principles (Green and Wolkenhauer 2013) and design principles (Green 2015).

For simplicity, let's "reset" our abstraction "tare" to the topological model, because the model driving the organizational design strategy—call it the design model—is abstract in all the ways the topological model is abstract and more besides. The simplification thereby focuses attention on ways in which the design model differs from the topological model—and, by extension, from the initial dynamical model.

Compared to the topological model, the design model contains more aggregations. For example, the design model represents CheY phosphorylation and CheB activation as "*k*-box output." This aggregation is, at the same time, a generalization and a structuration. For example, "*k*-box output" is a class, with instances biological as well as mechanical. The standard example of a mechanical instance is heater activation in a thermostat. The *k*-box representation is also a structuration, akin to the "A", "B," and "C" representations from the topological model. For the *k*-box represents whatever has such-and-such input and output (a position in a structure). (See Figure 4.)

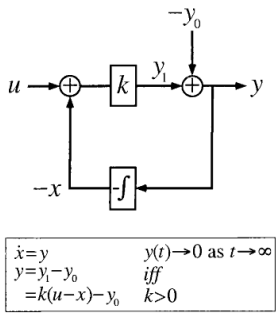


Fig. 2. A block diagram of integral feedback control. The variable u is the input for a process with gain k. The difference between the actual output y_1 and the steady-state output y_0 represents the normalized output or error, y. Integral control arises through the feedback loop in which the time integral of y, x, is fed back into the system. As a result, we have $\dot{x} = y$ and $y = 0$ at steady-state for all u. In the Barkai–Leibler model of the bacterial chemotaxis signaling system, the chemoattractant is the input, receptor activity is the output, and $-x$ approximates the methylation level of the receptors.

Figure 4 Organizational design for bacterial chemotaxis…and thermostats (Yi et al 2000).

The topological model is more abstract than the dynamical model, by virtue of containing various abstractions over protein identities. The design model, in turn, is more abstract than the topological and dynamical models, by virtue of also containing various abstractions over protein interactions. We can, therefore, arrange the various explanatory strategies along a continuum of abstraction type and severity. The dynamical modeling strategy, as our baseline, occupies the "low" end of our continuum. Next is topological analysis, which involves aggregations of and structurations from protein identities (or aggregations thereof). Then there is organizational design, which also involves aggregation of protein interactions as well as generalization and structuration of protein identities (or aggregations thereof).

5 Confirming the Analysis

I consider the foregoing to establish that each explanatory strategy invokes a different combination of abstraction types and that each targets its abstractions to different mechanistic details. Whether this result generalizes beyond my chosen case study awaits future research. There is some reason to expect an affirmative result. For if dynamical, topological, and design explanatory strategies differ as I claim—specifically, along dimensions of number and severity of generalizations and structurations—then we should expect the *more abstract* strategies to have *wider scope*. For the more general models likely have more instances, and the more structural models likely have more position occupants.

We find confirmation of this prediction for the case of robust perfect adaptation of *Bacillus subtilis* (*B.subtilis*) chemotaxis. Details of the organization design strategy for explaining why *E.coli* chemotaxis exhibits robust perfect adaptation *also* apply for explaining why *B.subtilis* chemotaxis exhibits robust perfect adaptation. But details of the corresponding dynamical mechanistic strategy do not. The organization design strategy, as we know, involves more generalization and structuration than the dynamical mechanistic strategy. This confirms our prediction.

Allow me to be brief with the details. Rao and Ordal (2007) develop a dynamic mechanistic explanation for the perfect robustness of chemotaxis for *B.subtilis*. Their explanatory strategy follows the same pattern as Barkai and Leibler's in the case of *E.coli*. But details differ. For example, according to Barkai and Leibler's model, CheB in *E.coli* demethylates only active receptor complexes; according to Rao and Ordal, CheB in *B.subtilis* demethylates inactive ones too. Again, according to Barkai and Leibler's model, without CheY *E.coli* runs but does not tumble; according to Rao and Ordal, without CheY *B.subtilis* tumbles but does not run. One more: according to Barkai and Leibler's model, *E.coli* without CheB cannot run; according to Rao and Ordal, *B.subtilis* without CheB can run. See Figure 5.

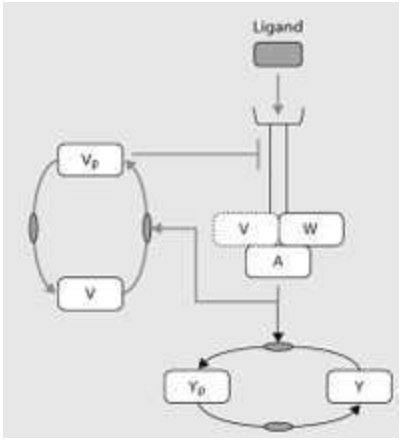


Figure 5: Chemotaxis network for *B.subtilis* (Rao and Ordal 2009).

So Barkai and Leibler's dynamical mechanistic explanation does not apply for the case of *B.subtilis*. But Yi and colleague's organizational design strategy does. For *B.subtilis*, like *E.coli*, exhibits robust perfect adaptation for chemotaxis if and only if it satisfies the characteristic equation for integral feedback control.

6 Toward Abstractive Mechanistic Explanation and its Affordances

Systems biological strategies for explaining the robust perfect adaptation of bacterial chemotaxis (in *E.coli*, *B.subtilis*, etc) apply mathematical techniques to network models. Dynamical, topological, and design strategies apply different techniques to explain the same phenomenon. Each explanatory strategy, moreover, applies its mathematical techniques to network models that embody different kinds and severities of these abstractions such as aggregations, generalizations, structurations. These abstraction types, accordingly, help to explain how these systems biological explanatory strategies differ from each other.

These abstraction types also provide a foundation for unifying various explanatory strategies from systems biology under the banner of mechanistic explanation. Let's consider well known kinds of mechanistic explanation as *standard*. Let's also follow Bechtel and Abrahamsen (2010) by considering *dynamical* mechanistic explanation as a mathematized species of standard mechanistic explanation.

Then let an *abstract network* be any network representation obtained by aggregating, generalizing, or structurating mechanistic details of the sort familiar in standard mechanistic explanation. Also let an *abstractive* mechanistic explanation be any explanation driven by applying mathematical techniques to an abstract network. See Figure 6.

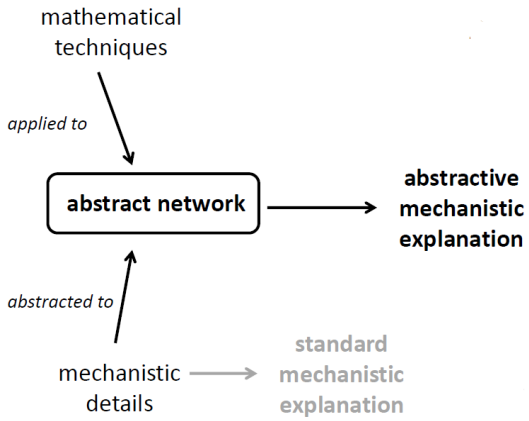


Figure 6. Relating standard and abstractive mechanistic explanation.

Then topological and organizational design explanatory strategies are mechanistic strategies—albeit abstractive ones. Topological explanations apply topological analysis

to aggregated and generalized mechanism networks. Organizational design explanations apply control systems engineering to aggregated, generalized, and structurated mechanism networks.

Both kinds of explanation are mechanistic, by virtue of being grounded upon mechanistic details. But both also provide explanatory affordances unavailable through standard mechanistic explanations, by virtue of being abstract. For example, by virtue of using generalizations, topological explanations should have a greater scope than their standard mechanistic counterparts. By virtue of using generalizations and structurations, organizational design explanations should have still greater scope.

That these abstractive mechanistic strategies use novel mathematical techniques is a side effect of their using novel abstractions (in comparison with standard mechanistic explanations and their dynamical cousins). These techniques, of course, support more general conclusions, with wider scope, than the kind of differential equation analysis available for dynamical mechanistic explanations. But the techniques do not explain why the strategies have broader scope.

References

U.Alon, M.G.Surette, N.Barkai, and S.Leibler, “Robustness in Bacterial Chemotaxis,” *Nature* 397 (2009), 168-171.

N.Barkai and S. Leibler. “Robustness in simple biochemical networks,” *Nature* 387 (1997), 913-917.

W.Bechtel and A.Abrahamsen, "Explanation: A Mechanistic Alternative," *Studies in History and Philosophy of Biological and Biomedical Science* 36 (2005), 421-441.

W.Bechtel and A.Abrahamsen, "Dynamic Mechanistic Explanation: Computational Modeling of Circadian Rhythms as an Exemplar for Cognitive Science," *Studies in History and Philosophy of Science* 41 (2010), 321-333.

H.C.Berg, *E.coli in Motion* (Springer, 2003).

P.A. Braillard, “Systems Biology and the Mechanistic Framework,” *History and Philosophy of the Life Sciences* 32.1 (2010), 43-62.

I.Brigandt, "Systems Biology and the Integration of Mechanistic Explanation and Mathematical Explanation," *Studies in History and Philosophy of Biological and Biomedical Sciences* 44 (2013), 477-492.

I.Brigandt, S.Green, and M.O’Malley, “Systems Biology and Mechanistic Explanation,” in S. Glennan and P. Illari (eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (*forthcoming*).

N.Cartwright, “Capacities and Abstraction,” in P. Kitcher and W. Salmon (eds.), *Scientific Explanation* (University of Minnesota Press, 1989), 349-356.

C.F.Craver, "When Mechanistic Models Explain," *Synthese* 153 (2006), 355-376.

S.Green, “Revisiting Generality in Biology: Systems Biology and the Quest for Design Principles,” *Biology and Philosophy* 30.5 (2015), 629-652.

S.Green and N.Jones, “Constraint-Based Reasoning for Search and Explanation: Strategies for Understanding Variation and Patterns in Biology,” *dialectica* 70.3 (2006), 343-374.

S.Green and O.Wolkenhauer, “Tracing Organizing Principles: Learning from the History of Systems Biology,” *History and Philosophy of the Life Sciences* 35 (2013), 553-576.

S. Haslanger, "What is a (Social) Structural Explanation?" *Philosophical Studies* 173 (2016), 113-130.

P.Huneman, "Topological Explanation and Robustness in Biological Sciences," *Synthese* 177 (2010), 213-245.

N.Jones, “Bowtie Structures, Pathway Diagrams, and Topological Explanation,” *Erkenntnis* 79.5 (2014), 1135-1155.

N.Jones and O.Wolkenhauer, “Diagrams as Locality Aids for Search and Explanation in Molecular Cell Biology,” *Biology and Philosophy* 27 (2012), 1135-1155.

M.Kollman, L.Løvdok, K.Bartholome, J.Timmer, and V.Sourjik, "Design Principles of a Bacterial Signalling Network," *Nature Letters* 438.24 (2005), 504-507.

A.Levy, “Three New Kinds of Mechanism,” *Biology and Philosophy* 28.1 (2013), 99-114.

A.Levy and W.Bechtel, “Abstraction and the Organization of Mechanisms,” *Philosophy of Science* 80 (2013), 241-261.

W.Ma, A.Trusina, H.El-Samad, W.A.Lim, and C.Tang, “Defining Network Topologies that Can Achieve Biochemical Adaptation,” *Cell* 138 (2009), 760-773.

P.Machamer, Lindley Darden, and C.F.Craver, "Thinking about Mechanisms," *Philosophy of Science* 67 (2000), 1-25.

D.Matthiessen, “Mechanistic Explanation in Systems Biology: Cellular Networks,” *British Journal for Philosophy of Science* (*forthcoming*).

D.J.Nicholson, “The Concept of Mechanism in Biology,” *Studies in History and Philosophy of Biologica and Biomedical Sciences* 43.1 (2012), 152-163.

S.A.G.Ordorica, “The Explanatory Role of Abstraction Processes in Models: The Case of Aggregations,” *Studies in History and Philosophy of Science* (2015), 161-167.

C. Pincock, “Abstract Explanations in Science,” *British Journal for the Philosophy of Science* 66 (2015), 857-878

J.Raerinne, “Robustness and Sensitivity of Biological Models,” *Philosophical Studies* 166 (2013), 285-303.

C.V.Rao and G.W.Ordal, “The Molecular Basis of Excitation and Adaptation during Chemotaxic Sensory Transduction in Bacteria,” in M. Collin and R. Schuch (eds.), *Bacterial Sensing and Signaling* (Karger: Basil, Switzerland, 2009), 33-64.

C.Rice, "Moving Beyond Causes: Optimality Models and Scientific Explanation," *Nous* 49 (2015), 589-615.

J.L.Spudich and D.E.Kochland, “Non-Genetic Individuality: Chance in the Single Cell,” *Nature* 262 (1976), 467-471.

G.H.Wadhams and J.P.Armitage, “Making Sense of It All: Bacterial Chemotaxis,” *Nature Reviews: Molecular Cell Biology* 5 (2004), 1024-1037.

A.G.Wouters, "Design Explanation: Determining the Constraints on What Can Be Alive," *Erkenntnis* 67 (2007), 65-80.

T.-M.Yi, Y.Huang, M.I.Simon, and J.Doyle, “Robust Perfect Adaptation in Bacterial Chemotaxis through Integral Feedback Control,” *PNAS* 97 (2000), 4649-4653.

How the Diachronic Theoretical Virtues Make an Epistemic Difference

Mike Keas • Professor of the History and Philosophy of Science • The College at Southwestern

Abstract. Among the virtues of good theories are those appropriately labeled diachronic: durability, fruitfulness, and applicability—the last of which is insufficiently recognized. Diachronic theoretical virtues *cannot* be instantiated in the original construction of a theory; subsequent development is required. By contrast, one *can* assess the degree to which a theory exhibits the following nine non-diachronic theoretical virtues in a theory's original construction: evidential accuracy, causal adequacy, explanatory depth, internal consistency, internal coherence, universal coherence, beauty, simplicity, and unification. The distinction between diachronic and non-diachronic virtues is important for understanding the role and epistemic standing of each theoretical virtue.

Keywords. Theoretical virtues, durability, fruitfulness, prediction, and science-technology relations.

1. Introduction. Theoretical virtues are the traits of a theory that show it is probably true or worth accepting. Although the identification, characterization, classification, and epistemic standing of theory virtues are debated by philosophers and by participants in specific theoretical disputes, many scholars agree that these virtues help us to infer which rival theory is the best explanation (Lipton 2004). The most widely accepted theories across the disciplines usually exhibit many of the same theoretical virtues listed below. Each virtue class contains at least three virtues that sequentially follow a repeating pattern of progressive disclosure or expansion. In another forthcoming essay (Keas 2017) I argue for this new systematization of the theoretical virtues. In the present essay I focus on the diachronic class of virtues in contrast with the non-diachronic virtues. One can assess the degree to which a theory exhibits the non-diachronic virtues from the time a theory is initially framed. However, no theory, in its original construction, can instantiate

the diachronic virtues: durability, fruitfulness, or applicability. These virtues are instantiable only as a theory is later refined or applied.

Evidential virtues
1. Evidential accuracy: A theory (T) fits the empirical evidence well (regardless of causal claims).
2. Causal adequacy: T's causal factors plausibly produce the effects (evidence) in need of explanation.
3. Explanatory depth: T excels in causal history depth or in other depth measures such as the range of counterfactual questions that its law-like generalizations answer regarding the item being explained.

Coherential virtues
4. Internal consistency: T's components are related to each other logically.
5. Internal coherence: T's components are coordinated into an intuitively plausible whole; T lacks ad hoc hypotheses—theoretical components merely tacked on to solve isolated problems.
6. Universal coherence: T sits well with (or is not obviously contrary to) other warranted beliefs.

Aesthetic virtues
7. Beauty: T evokes aesthetic pleasure in properly functioning and sufficiently informed persons.
8. Simplicity: T explains the *same facts* as rivals, but with *less* theoretical content.
9. Unification: T explains *more kinds of facts* than rivals with the *same* amount of theoretical content.

Diachronic virtues
10. Durability: T has survived testing by successful prediction or plausible accommodation of new data.
11. Fruitfulness: T has generated additional discovery by means such as successful novel prediction, unification, and non ad hoc theoretical elaboration.
12. Applicability: T has guided strategic action or control, such as in science-based technology.

We will survey the first nine virtues only to the brief extent needed to recognize how one can assess the degree to which a theory exhibits these theoretical virtues in its original construction. This will, by contrast, enable us to appreciate the unique temporal character of the diachronic theoretical virtues.

2. Non-Diachronic Theoretical Virtues. We begin with the first three virtues. *Evidential accuracy*, which is how well a theory fits the relevant data, can be assessed from the theory's original construction. Often a theory will also, from its inception, specify *causally adequate* mechanisms to produce the phenomena in question. Such is not necessarily the case, as Alfred Wegener's theory of continental drift illustrates. His theory enjoyed considerable evidential accuracy despite its lack of a plausible cause to move the continents. *Explanatory depth* is also instantiated in a theory's initial formulation if, for example, the

theory answers a large range of counterfactual questions about a kind of phenomenon using the resources of its law-like generalizations.

The remaining six non-diachronic theoretical virtues likewise can be exhibited in the initial formation of a theory. A theory may be constructed in a logical manner so as to produce *internal consistency*. Beyond that, the theoretical components might be well coordinated into an intuitively plausible whole (avoiding ad hoc hypotheses), thus generating the theoretical virtue of *internal coherence*. If the theory sits well with (or is not obviously contrary to) other warranted beliefs, then it possesses the virtue of *universal coherence*. A new theory might even evoke aesthetic pleasure in the minds of experts, which constitutes theoretical *beauty*. The closely related virtues of simplicity and unification also might be instantiated in the initial formation of a theory: explaining the same facts as rival theories but with less theoretical content (*simplicity*), and explaining more kinds of facts than rivals with the same amount of theoretical content (*unification*).

Much more could be said about the first nine virtues outlined above (Keas 2017), but this is sufficient to recognize them as a group of theoretical virtues that can, in principle, be instantiated in a theory's original formation. This common trait remains characteristic of these virtues even (largely) under the disparate accounts found in the literature of how to characterize each virtue. Let us now explore the chief diachronic theoretical virtues in contrast to the non-diachronic virtues.

3. Diachronic Theoretical Virtues. Durability, fruitfulness, and applicability, which I recognize as the chief diachronic theoretical virtues, can only be instantiated as a theory is cultivated *after* its origin. This necessarily extended temporal dimension of the diachronic virtues is, arguably, of considerable epistemic importance. But even if one endorses the arguments that discount the epistemic significance of this temporal component (Mayo 2014), one still should acknowledge a group of virtues that (unlike the other the-

oretical virtues) can only be instantiated in a theory *after* its initial formulation. Time is of their essence in a manner that goes beyond the trivial truth that all human endeavor is temporal. McMullin (2014) has lead the way in articulating the epistemic significance of two of the three main diachronic virtues: durability and fruitfulness (I recognize McMullin's third diachronic virtue of "consilience" as a mode of fruitfulness). Applicability, largely overlooked as a theory virtue, is another important member of this diachronic category, as I shall demonstrate.

3.1. Durability. Durability, a virtue term McMullin (2014) recommended, refers to the favorable epistemic condition of a theory that has survived testing by successful prediction or by plausible accommodation of new unanticipated data (or both). Popular or long-lived theories are not necessarily durable in the epistemic sense in view here. Equating durability with popularity or tradition is fallacious. While testability is a pragmatically admirable trait of a theory, it is not an intrinsic epistemic characteristic of a theory; many testable theories have failed too many tests to be acceptable. Steel (2010, 18) notes that the "more precise and informative a theory's empirical predictions are, the greater its testability." The more testable a theory is, the more durable it would prove itself to be if it passes the tests. A theory that scores low in testability has little potential to exhibit durability.

Despite the leading role of predictive success in many areas of science, it is less prominent in some reputable scientific theories that are, nevertheless, well endowed with other virtues. Successful prediction is very frequently part of explaining "how things work," but less routine in explaining "how things originated"—as in theories about the history of the cosmos, earth, and life (Cleland 2011, but Winther 2009 argues otherwise). Successful historical theories typically enjoy other forms of durability, most notably a track record of plausible accommodation of new data that, although not predicted, came to light after the theory's origin. The durability of a theory suffers if one or more of its predictions are disconfirmed

or when theorists respond to disconfirming evidence by modifying the theory with ad hoc hypotheses—theoretical components merely "tacked on" to solve isolated problems. Although initially a theory may exhibit a high degree of evidential accuracy (or any other of the first nine virtues in my systematization), it is impossible for a newborn theory to instantiate the virtue of durability—this *takes time* in a sense not required by the non-diachronic virtues. A similar necessary temporal dimension characterizes fruitfulness.

3.2. Fruitfulness. Fruitfulness, also known as fertility or fecundity, is another diachronic theoretical virtue. A theory is fruitful if, over time, it generates additional discovery by means such as successful novel prediction, unification, and non ad hoc theoretical elaboration. While durability is about conservation (a theory passing tests to survive), fruitfulness is about innovation (a theory stimulating further discovery). When a prediction formulated in the context of a theory's construction is later verified, this successful predictive outcome increases the virtue of durability in that theory. By contrast, a *novel* prediction is one that was not conceived in conjunction with a theory's construction, but that nevertheless follows reasonably from it. When such a novel prediction is confirmed by observation, a theory exhibits more fruitfulness.

The closely related diachronic character of durability and fruitfulness is well illustrated in the discovery of the first two planets beyond Saturn. Soon after Friedrich William Herschel unexpectedly discovered Uranus in 1781, astronomers noted that its observed motion strayed from what contemporary Newtonian mechanics predicted of such a planet. However, given the overall theoretically virtuous status of Newtonian physics up through that time (including its durability due to its success in testing), most astronomers expected a forthcoming way to make Uranus compliant with established theory. Even rejecting the anomalous data as "inaccurate" seemed reasonable early on. By the 1830s, however, the possibility of a perturbing planet beyond Uranus became a more reasonable and popular speculation, despite the ab-

sence of a precise novel prediction of where to find such a planet. By this time many astronomers were modestly confident in the accumulated data of Uranus' positions in the sky.

This brings us to the celebrated successful novel prediction of 1845-1846. Based principally on Newtonian physics and the well-known irregularities in Uranus' motion, two astronomers independently predicted where another unknown perturbing planet (later called Neptune) was likely located. Le Verrier's estimate of the planet's location was the most accurate (correct within one degree), as confirmed by a German astronomer on September 23, 1846. The (*fruitful*) novel prediction of Neptune was born within the context of a *durable* Newtonian orbital mechanics research tradition and the unexpected discovery of Uranus with its anomalous motions. The sensational success of this novel prediction (the discovery of Neptune) also rendered Uranus a Newtonian-compliant planet—thus further vindicating earlier provisional toleration of Uranus' anomalies, a toleration that had been justified by yet earlier Newtonian durability and fruitfulness.

Smith's (2010; 2014) landmark study of gravity theory from Newton to the present further illuminates the durability and fruitfulness of this research tradition, and it includes the case histories of Uranus and Neptune. Smith was surprised that the principal kind of question being tested was not "Do the calculated motions [e.g., of Uranus] agree with the observed motions?" Rather it was: "Can robust physical sources compatible with Newtonian theory be found for each clear, systematic discrepancy between the calculated and the observed motions?" Neptune (as novelly predicted) turned out to be such a robust physical source. However scientists failed over a half century to find a robust (detectable) physical source for the Newtonian-defying behavior of Mercury—a tiny anomaly in the precession of its perihelion. But this failure, which Einstein solved by way of theory replacement, does not completely diminish the enduring epistemic significance of two centuries of Newtonian durability and fruitfulness, as Hanson (1962) inaccurately suggested. Smith notes: "All the other discrepancies ended up revealing some detail of our plane-

tary system, the least subtle of which was Neptune, that theretofore had not been taken into account in the calculations" (2010, 552).

Such serial Newtonian problem solving became (almost always) ever more empirically constrained in a spiral of upward progress. For example, Uranus' temporarily Newtonian-defying behavior "would have been masked if the significantly larger gravitational effects of Saturn on Uranus had not been included in the calculation first." Smith explains further:

> So, the discovery of Neptune provided evidence not only for Newton's theory, but also for the specific aspects of Saturn that entered into calculating its effects on Uranus, for these were no less presupposed in the anomaly that emerged than Newton's theory was. The point generalizes. Each time a discrepancy emerges and a robust physical source for it is found, that source is incorporated into the new calculations, and the process is repeated, typically with still smaller discrepancies emerging that were often theretofore masked in the calculations. So, what was being tested each time when a new discrepancy emerged and a physical source for it was being sought was not only Newtonian theory, but also all the previously identified details that make a difference and the differences they were said to make without which the further systematic discrepancy would not have emerged. (2010, 552-53)

On display is an interlocking of durability (passing tests to survive) and fruitfulness (stimulating further discovery) that is supportive of scientific realism. "This shows that increasingly strong evidence was accruing to Newtonian theory over the first two hundred years of orbital research based on it," Smith concludes. This point (with some qualification) extends even to Einstein's theoretical innovation that was partly justified by the unruly perihelion of Mercury. Einstein's achievement was, to some degree, a continuation of this same progressive spiral, as Smith deftly explains:

> As is well known, Einstein required Newtonian gravitation to hold in an asymptotic limit as he developed his new theory of gravity—specifically in a static, weak-field limit. That he did so was just as well because the 43 arc-seconds per century anomaly in the perihelion of Mercury that was initially the sole evidence for his theory presupposes Newtonian gravity…. As a matter of historical fact, all of the details singled out as making detectable differences during the two centuries of prior research carried over intact into post-Einstein orbital mechanics. *Save for some qualifications concerning levels of precision, the same details are still making the same differences as before.*... So, Newtonian theory must still have some sort of claim to being knowledge. (2010, 556-57)

Smith's continuity-of-knowledge claim invites comment. While much of the metaphysics associated with Newtonian theory has been repudiated, we nevertheless see an impressive degree of fruitful scientific continuity from Newtonian to modern physics (at least in the particular ways that Smith documents). In sum, Newtonian orbital mechanics enjoyed increasingly impressive interlocking durability and fruitfulness over multiple centuries, and its approximate legitimacy (not counting discarded Newtonian metaphysics) remains similarly well-grounded today under the revisionary umbrella of modern physics.

Though some philosophers have argued to the contrary (Collins 1994; Harker 2008), many scientists and philosophers think that predictive success—especially novel predictive success—is a stronger indicator of likely approximate truth than a theory's accommodation of data (Douglas and Magnus 2013). According to my systematization (which illuminates but does not settle this thorny issue), data accommodation refers to a theory's initial instantiation of the evidential virtues (evidential accuracy, causal adequacy, and explanatory depth), and a theory's subsequent instantiation of certain diachronic virtues, namely non-predictive durability (plausibly making sense of new unanticipated data) and non-predictive fruitfulness (especially non ad hoc theoretical elaboration that makes sense of new unanticipated data).

3.2.1 Unification as a Mode of Fruitfulness. Fruitful theory elaboration, whether by means of successful novel prediction or non ad hoc theoretical elaboration that makes sense of unanticipated evidence, often also makes sense of *new kinds* of data, and thus is additionally recognized as increasing a theory's unification. Earlier we encountered unification as a non-diachronic (aesthetic) theoretical virtue. The diachronic increase of unification differs somewhat from its non-diachronic cousin. The historian and philosopher of science William Whewell (1794–1866) called diachronic unification "consilience." When a theory explains a new domain of facts in a surprising way, then it is fruitful in a consilient manner. McMullin writes in this regard:

> A good theory will often display remarkable powers of unification, making different classes of phenomena "leap together" over the course of time. Domains previously thought to be disparate now become one, the textbook example, of course, being Maxwell's unification of magnetism, electricity, and light. Examples abound in recent science, a particularly striking one being the development of the plate-tectonic model in geology. Assuming that this unifying power manifests itself over time, it testifies to the epistemic resources of the original theory and hence to that theory's having been more than mere accommodation. (2014, 505)

McMullin contrasts diachronic unification with its non-diachronic counterpart: "If the unification was achieved by the original theory, however, the virtue involved would no longer be diachronic." Instead, it would count (in my systematization) as an aesthetic theoretical virtue that I simply call "unification," and that Lipton calls "variety" (and yet others call "broad scope"). Lipton favors the assumption that such "heterogeneous evidence provides more support than the same amount of very similar evidence" (Lipton 2004, 168). Despite my own inclination to accept Lipton's point, I recognize this as a somewhat debatable assumption about the epistemic significance of an aesthetic property. However, when unification increases

over time, especially by means of surprising convergences, then unification is less likely the result of the idiosyncratic aesthetic predispositions and clever accommodating skills of a theorist during theory formation. Thus fruitful diachronic unification has greater confirmatory power than a theory's initial degree of aesthetic unification.

3.2.2 The Role of Prediction in the Diachronic Virtues. Drawing from Douglas' work on the relationship of prediction to inferring the best explanation, I argue that predictive success (in the first two diachronic virtues explored above) extends the epistemic work of many non-diachronic theoretical virtues such as causal adequacy, explanatory depth, beauty, simplicity, and unification. These latter theory traits, which she collectively labels as "explanatory,"

> appeal to us, not just because we are aesthetically driven creatures but because such virtues help us to use the explanation to think and, in particular, to think our way through to new predictions, new tests, new rigors for our beautiful explanation. (2009, 460)

Douglas also notes:

> Predictions are valuable because they force us (when followed through) to test our theories, because they have the potential to expand our knowledge into new realms and because they hold out the possibility (if successful) of gaining some measure of control over natural processes. (2009, 455)

Transposing Douglas' insights into my taxonomic terms, predictions are valuable because they figure into all three of the major diachronic virtues: durability (testing theories successfully), fruitfulness (expanding "our knowledge into new realms"), and applicability (which includes "gaining some measure of control over natural processes"). Moreover, the operation of prediction ("saying before" at least in a logical if not

temporal sense) in these three theoretical virtues further supports my classification of them as diachronic. Lets us now explore the last major diachronic virtue of applicability.

3.3. Applicability. Applicability refers to when a theory is used to guide successful action (e.g., prepare for a natural disaster) or to enhance technological control (e.g., genetic engineering). High degrees of the virtue of applicability obtain when a theory that is used to guide such action or control provides more effective outcomes than what is possible in the absence of the theory. Successful scientific theories constitute *knowledge* of the world (knowing *that*), not *control* over the world (which is mainly knowing *how*) for practical (non-theoretical) purposes. In this regard Strevens (2008, 3) notes: "If science provides anything of intrinsic value, it is explanation. Prediction and control are useful … but when science is pursued as an end rather than as a means, it is for the sake of understanding." But even after the intrinsic good of a theoretically virtuous explanation is in hand, one of several possible additional confirmatory diachronic (predictive or controlling) virtues might be acquired by a theory, including applicability. In such cases a good theory just gets better—even more confidence in its probable truth is justified.

Although scientific experiments use technological control, they do so to test scientific theories—so the main function is still to understand nature, not to control it. However, especially in the case of theories supported by experimentally verified prediction, such foreknowledge and laboratory control might be exploited to achieve practical aims such as device fabrication or medical intervention. But in any case, one cannot *apply* scientific knowledge until *after* one first *obtains* it. This necessary time lapse makes applicability diachronic.

To obtain scientific knowledge we search for a theory that (initially) exhibits many of the non-diachronic theoretical virtues. Subsequent work aimed at theory testing and elaboration might produce the additionally confirming presence of the diachronic virtues of durability and fruitfulness. At some point in

this dance of virtue-driven theory assessment and refinement, sufficient confidence in a particular theory might spur attempts to apply it as the basis for a new or improved technology. If the derived science-based technology actually works, then the "applied theory" has acquired the additional theoretical virtue of applicability. Because this requires additional time after initial theory formation, the diachronic classification of applicability is appropriate.

Although the application of scientific theories constitutes one aspect of technology, most of technology involves the empirical discovery of "know how" knowledge without crucially presupposing or immediately applying any particular scientific theory. Indeed, the relation between science and technology is not a simple one-way linear affair (Radder 2009; Douglas 2014). But this "emancipation" of technology from subordination to science, accomplished by historians and philosophers of technology between 1960 and 1990 (Houkes 2009, 310), should not obscure the epistemic significance of instances of technological innovation made possible, in part, by *applied* scientific theory.

This point is in harmony with the so-called demise of the "pure vs. applied science" dichotomy. Understanding and controlling nature are closely related, as our study of the diachronic theoretical virtues, including applicability, indicates. Douglas (2014, 62) surfaces some of the subtlety of this argument when, on the one hand, she proclaims: "With the pure vs. applied distinction removed, scientific progress can be defined in terms of the increased capacity to predict, control, manipulate, and intervene in various contexts." But then, on the other hand, in a footnote she recoils partially: "To be clear, while I think this is a useful rubric for scientific progress, it is not a remotely sufficient account for how one should assess scientific theories." Other (non-diachronic) theoretical virtues that are complementary to, but less weighty epistemically than, prediction and control also play important roles in theory assessment, she suggests. Consideration of the nine major non-diachronic theoretical virtues systematized in Sections 1 and 2 drives this point home.

How exactly is applicability a diachronic theory trait that is *epistemic* (helping to indicate likely truth) in view of the obvious *pragmatic* orientation of technological application? Agazzi observes that some technological projects "are designed or projected in advance, as the concrete application of knowledge provided by a given science or set of sciences" (Agazzi 2014, 308). If a project of this kind actually works as predicted, then this reinforces our confidence in the theory base that helped guide such action in the world. Agazzi further notes:

> The predictions 'contained' in the project actually are the predictions made by the scientific theories which have permitted the proposal of the complex *noema* that constitutes the project, and contains not only prescriptions as to the way of realising the structure of the machine but also as to its functioning. This functioning is something that happens; it is a state of affairs that constitutes a confirmation of the theories used in projecting the machine. (309)

Although Agazzi's scientific realism overstates the epistemic reach of applicability, it is helpful nonetheless as a corrective to other philosophical errors:

> A mature science is a science that has given rise to a significant technology. This means, for example, that we can provisionally admit certain theories that are 'empirically adequate,' without admitting their truth as van Fraassen says, until we have significant predictions confirming them. This fact (especially in conjunction with other 'virtues' discussed in the literature) already justifies attributing truth and ontological reference to them, but the existence of technological applications is the last decisive step that assures that they have been able to adequately treat those aspects of reality they intended to treat. These last words are very important. They underline the fact that technological success does not eliminate the partial or limited scope of scientific theories. The fact that we can use classical mechanics in creating many machines or for sending rockets into space certainly means that this mechanics is true of its

> objects and therefore 'tells a true story' about certain aspects of reality. This can also be expressed by saying that this theory is partially true of reality, but only if we mean that it does not speak about the totality of the attributes of reality, and that, consequently, it can speak properly only of such referents that possess these attributes. In other words, it is not correct to say that this mechanics is true regarding the whole of reality because other aspects of reality exist that must be accounted for by means of other theories which, in turn, can be used as a basis for different technologies. (310-11)

To nuance Agazzi's insightful but somewhat inflated epistemic role for applicability, we can observe that this theoretical virtue is not commonly operative in certain scientific domains. For example, scientific theories of "how things originated" (history of nature) lead to fewer technological applications than scientific theories of "how things work." Part of the reason for the infrequent applicability of origins theories is the smaller role that experimentally controlled prediction plays in such theorization. For example, much of the data that allows us to reconstruct the *history* of earth's surface is collected by means of passive field observations, rather than by laboratory experiments that make precise predictions and technological control more feasible.

4. Conclusion. The diachronic theoretical virtues possess a temporal dimension that is absent from the other theoretical virtues. They can only be instantiated *after* a theory's initial formulation—when it has had opportunity to be tested, elaborated, and applied. Durability, fruitfulness, and applicability build upon the initial theory assessment process governed by the non-diachronic virtues (the evidential, coherential, and aesthetic theoretical virtues). The cumulative result, when successful, is a mature theory with an even greater probability of being true than an infant theory that has not yet had the opportunity to show whether it will possess the diachronic theoretical virtues (anti-realists are invited to interject their own alternative

to this realist understanding of the theoretical virtues). So, the distinction between diachronic and non-diachronic virtues is important for an adequate account of theory evaluation.

The three major diachronic theoretical virtues are also better understood when they are recognized as related to each other in the following progressive sequence. Durability is instantiated as a theory passes more rigorous tests in a series of encounters with the world, especially by successful prediction and plausible accommodation of new evidence. Fruitfulness discloses a theory's resourcefulness yet further through innovation—stimulating additional discovery by successful novel prediction, unification, non ad hoc theoretical elaboration, and other means. At last, applicability expands the epistemic accountability of a theory into the final frontier: the vast domain of practical action. This virtue is instantiated when a theory helps us to interact with the world successfully, most notably by technological control. Together, these diachronic theoretical virtues provide an ongoing and epistemically intensified means of theory development that complements the non-diachronic virtue assessment process that begins in a theory's original construction.

Applicability, *as a theoretical virtue*, has not received the attention it deserves. Surprisingly, it is absent from every theoretical virtue list I have encountered. My work sketches a way to understand applicability in relation to the other diachronic virtues, and the larger group of non-diachronic virtues. This endeavor promises to illuminate, among other things, discussion of realism vs. anti-realism, science-technology relations, and inference to the best explanation.

References

Agazzi, Evandro. 2014. *Scientific Objectivity and Its Contexts*. Cham: Springer.

Cleland, Carol E. 2011. "Prediction and Explanation in Historical Natural Science." *British Journal for the Philosophy of Science* 62 (3):551-582.

Collins, Robin. 1994. "Against the Epistemic Value of Prediction over Accommodation." *Noûs* 28 (2):210-224.

Douglas, Heather E. 2009. "Reintroducing Prediction to Explanation." *Philosophy of Science* 76 (4):444-463.

——— 2013. "The Value of Cognitive Values." *Philosophy of Science* 80 (5):796-806.

——— 2014. "Pure Science and the Problem of Progress." *Studies in History and Philosophy of Science Part A* 46:55-63.

Douglas, Heather, and P. D. Magnus. 2013. "State of the Field: Why Novel Prediction Matters." *Studies in History and Philosophy of Science Part A* 44 (4):580-589.

Hanson, Norwood Russell. 1962. "Leverrier: The Zenith and Nadir of Newtonian Mechanics." *Isis* 53 (3):359-378.

Harker, David. 2008. "On the Predilections for Predictions." *The British Journal for the Philosophy of Science* 59 (3):429-453.

Houkes, Wybo. 2009. "The Nature of Technological Knowledge." In *Philosophy of Technology and Engineering Sciences*, 309-350. Amsterdam: North-Holland.

Lipton, Peter. 2004. *Inference to the Best Explanation*. 2nd ed. London: Routledge.

Mayo, D. 2014. "Some Surprising Facts About (the Problem of) Surprising Facts." *Studies in History and Philosophy of Science Part A* 45:79-86.

McMullin, Ernan. 2014. "The Virtues of a Good Theory." In *The Routledge Companion to Philosophy of Science*, ed. Martin Curd and Stathis Psillos, 561-571. New York: Routledge.

Radder, Hans. 2009. "Science, Technology and the Science-Technology Relationship." In *Philosophy of Technology and Engineering Sciences*, ed. A. Meijers, 65-91. Amsterdam: North Holland.

Smith, George E. 2010. "Revisiting Accepted Science: The Indispensability of the History of Science." *Monist* 93 (4):545-579.

——— 2014. "Closing the Loop: Testing Newtonian Gravity, Then and Now." In *Newton and Empiricism*, ed. Zvi Biener and Eric Schliesser, 262-351. New York: Oxford University Press.

Steel, Daniel. 2010. "Epistemic Values and the Argument from Inductive Risk." *Philosophy of Science* 77 (1):14-34.

Strevens, Michael. 2008. *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Winther, R. G. (2009). Prediction in selectionist evolutionary theory. *Philosophy of Science, 76*(5), 889-901.

PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association -288-

Reconciling axiomatic quantum field theory with cutoff-dependent particle physics

Adam Koberinski[1]

[1]Department of Philosophy, Western University

Abstract

The debate between Fraser and Wallace (2011) over the foundations of quantum field theory (QFT) has spawned increased focus on both the axiomatic and conventional formalisms. The debate has set the tone for future foundational analysis, and has forced philosophers to "pick a side". The two are seen as competing research programs, and the major divide between the two manifests in how each handles renormalization. In this paper I argue that the terms set by the Fraser-Wallace debate are misleading. AQFT and CQFT should be viewed as complementary formalisms that start from the same physical basis. Further, the focus on cutoffs as demarcating the two approaches is also highly misleading. Though their methods differ, both axiomatic and conventional QFT seek to use the same physical principles to explain the same domain of phenomena.

1 Introduction

Foundational invesitgation into quantum field theory (QFT) has emerged as a flourishing enterprise in philosophy of science, thanks largely to work done in axiomatic QFT (AQFT), particularly the C^*-algebraic approach enocoded by the Haag-Kastler axioms (Haag and Kastler 1964). Despite the methodological disconnect with 'conventional' approaches to QFT (CQFT), AQFT has been defended by Fraser (2009) as supplying a firmer foundation from which to conduct philosophical analyses. Though this is one of few explicit defenses of AQFT, the widespread use of algebraic methods in philosophical literature on QFT would lead one to believe that Fraser is merely making explicit the assumptions in her field. Recently, Wallace (2006; 2011) has questioned the focus on AQFT, arguing that CQFT is the better candidate for analysis. Since CQFT is the theory that has been emprically successful—the Standard Model of particle physics is built from CQFTs—and AQFT has yet to reproduce these results, Wallace argues that we should focus analysis on CQFT rather than AQFT. Fraser's (2011) reply has set up what is now known as the Fraser-Wallace debate over the foundations of QFT. The debate has set the tone for future foundational analysis, and seems to force philosophers to "pick a side"—you either work in AQFT or CQFT. The two are seen as competing research programs, and the major divide between the two manifests in how each handles renormalization. AQFT requires strict Poincaré covariance at arbitrarily small length scales, while the renormalization group (RG) methods in CQFT allow for a small-scale cutoff, below which QFTs needn't be well-defined.

In this paper I argue that the terms set by the Fraser-Wallace debate are misleading. One needn't view AQFT and CQFT as rival research programs; in fact, this view is

detrimental to understanding the history and methodology of QFT. AQFT and CQFT should be viewed as complementary formalisms that start from the same physical basis. Further, the focus on cutoffs as demarcating the two approaches is also highly misleading: AQFT can accommodate cutoffs and RG methods, and CQFT does not explicitly require cutoffs. The focus on cutoffs as essential to CQFT could mistakenly be taken to mean that CQFT depends on cutoffs actually *being physical*, in the same way that cutoffs are physical in condensed matter physics (CMP). I will argue that this is not the case: cutoffs needn't be physical in any sense. Even if cutoffs are *physically significant*, that does not entail that the cutoffs are themselves physical. Specifically, RG methods provide no principled grounds for thinking that cutoffs are "real" in the sense of signifying a breakdown of field theories generally. Since Wallace (2011) set the terms of the debate, the bulk of the arguments in this paper will be in reference to that paper. I do not claim that Wallace holds all (or even most) of the views against which I argue; rather, I use his paper to clarify potential misconceptions that could arise from the debate. Renormalization is not central to the physical content of QFT, and the different ways of handling renormalization do not mark AQFT and CQFT as different research programs. We should instead view the formalisms as complementary: though their methods differ, both seek to use the same physical principles to explain the same domain of phenomena.

2 Renormalization and the relationship between AQFT and CQFT

Wallace (2011) emphasizes the ineliminable dependence on cutoffs in CQFT, along with the success of RG methods for providing a physical motivation for cutoffs, as the wedge which drives AQFT and CQFT apart. For Wallace, AQFT cannot deal with physical cutoffs. Since RG methods have physically legitimized cutoffs, AQFT and CQFT have differing physical content and must therefore be considered a different research program (2011, Sec. 2). I disagree with this characterization on two fronts. First, AQFT has the resources to incorporate RG methods when needed. Though typical axioms make no metion of scaling behaviour, even the most rigid of axiomatic approaches—algebraic QFT as codified in the Haag-Kastler axioms—can incorporate something like RG flows.[1] Second, the calculational dependence on cutoffs in CQFT may not signal the physical existence of cutoffs.

So, are cutoffs really that problematic for AQFT? Many axiomatic approaches to QFT make no recourse to cutoffs, either explicitly or implicitly. An explicit forbidding of cutoffs would mean that one of the axioms/postulates of the theory claimed that the theory is empirically adequate at all spacetime length scales. Even if any axiomatization contained such an axiom (none do), it would be hard to imagine what sort of work it would do in derivations. Presumably, such a system could be modified to remove the guilty axiom, without spoiling any physically useful theorems. One should therefore not be concerned with an explicit ban on cutoffs in AQFT.

The more interesting case is when cutoffs are implicitly rejected by a particular theory.

[1] See Buchholz and Verch (1995) for an example of scaling algebras playing the role of RG flows.

There are two common assumptions in AQFT that are problematic for handling cutoffs: strongly continuous implementations of Lorentz invariance, and the association of algebras with arbitrarily small open bounded regions of spacetime. Though the latter is not common to all axiomatic QFTs (the Wightman axioms deal directly with quantum fields, rather than algebras), the dominant axiomatization in terms of C^* algebras—the Haag-Kastler axioms—define QFTs in terms of algebras of observables corresponding to open, bounded regions of spacetime.[2] It is implicit that for any open bounded spacetime region, *no matter how small*, one can define an algebra of observables satisfying the other axioms defining QFT. If cutoffs are physical, one might conclude that there should be a principled limit to the size of regions on which we can define algebras corresponding to observables in QFT. If the cutoff scale is physically relevant, and only CQFT predicts its existence, we might be tempted to conclude that the two are different, competing theories. However, there are several possibilities for reconciling AQFT and cutoffs, which I will outline below. These remedies are largely independent of one another, and organized in terms of increasing foundational disagreement with Wallace's view of cutoffs. The "quick fixes" proposed first lead to further conceptual worries, and I therefore endorse the option in Sec. 2.3, which is the biggest departure from taking cutoffs as physical in CQFT. Nevertheless, all the options sketched below are more-or-less viable. Section 2.4 outlines reasons for thinking that *both* AQFT and CQFT suffer the same conceptual challenges if cutoffs *really are physical*.

[2]Since algebraic QFT is prima facie the most problematic, I will deal primarily with algebraic QFT in this paper. The reader can take AQFT to stand for axiomatic QFT or algebraic QFT for the remainder of this paper. The reader should also note that constructive QFT is another important strand of rigorous QFT. Though it is conceptually distinct from AQFT, the two projects often overlap.

2.1 Possibilities for cutoffs in AQFT

Just because we need to associate an algebra with any arbitrary open bounded region of spacetime, we are not therefore compelled to make this algebra interesting. One way that cutoffs could be introduced into AQFT is to specify that regions smaller than some 4-volume Λ are to be uniformly assigned trivial algebras, i.e., algebras containing only multiples of the identity. Such assignments would be consistent with the demand that all open bounded regions of spacetime be assigned an algebra, but it would make the cutoff physically relevant, since no information about local parameters would be contained in regions smaller than Λ.

Though this solution is available, it is admittedly somewhat ad hoc. Even worse, it violates one of the crucial Haag-Kastler axioms: that of weak additivity. The axiom of weak additivity states that, for *every* closed, bounded region $\mathcal{O}$ of Minkowski spacetime $\mathcal{M}$, the C^* norm closure of the algebras $\mathfrak{A}(\mathcal{O}+\alpha)$ for $\alpha \in \mathbb{R}^4$ is just the quasilocal algebra for the whole spacetime, $\mathfrak{A}(\mathcal{M})$.[3] There are two reasons why this is a problem for introducing cutoffs in the way described above. First, we run into the problem that the quasilocal algebra corresponding to the whole of $\mathcal{M}$ can be constructed from *any* algebra corresponding to *any* closed, bounded region $\mathcal{O}$. The norm closure of extensions of a trivial algebra will not produce any interesting algebra as a result, so regions smaller than the cutoff Λ will violate weak additivity. Second, extensions of an arbitrary region $\mathcal{O}$ by some $\alpha < \Lambda$ should not be physical if Minkowski spacetime breaks down at scales below Λ. In the spirit of the first ad hoc axiom modification, weak additivity could be modified to exclude regions $\mathcal{O}_{small} < \Lambda$, and arbitrary extensions $\alpha_{small} < \Lambda$. However,

[3]See Ruetsche (2011), especially chapters 4 and 5 for an introduction to algebraic QFT. For a more comprehensive review of algebraic QFT, see Halvorson and Müger (2007).

there seems to be no principled reason for choosing a specific value of Λ, and one may question the naturalness of such axioms. This makes the solution of simple axiom modification less tempting, and forces us to admit that AQFT—at least in its current guise—is in conflict with approaches to QFT that take cutoffs as physically meaningful, since the basic axioms are currently in direct conflict with the introduction of cutoffs. If we admit that there is currently no room in the formalism of AQFT for cutoffs, are we doomed to take AQFT as (incorrectly) positing its own validity at all energy scales?

2.2 No cutoffs? No problem

If QFT methods are only applicable up to some cutoff energy, and we expect QFT to incorporate this fact, we are saying that a good theory should signal its own demise. The formal necessity of cutoffs in the formalism of CQFT has lead to the idea that our best theories will continue to be an increasing hierarchy of effective field theories. Each field theory requires cutoffs to be implemented at a certain energy scale, and this signals the field theory's domain of applicability. If supplanted by a successor field theory, one expects that the new theory's low energy regime reduces to the old theory, and further that the new theory will itself have a higher energy cutoff. Following this approach, the conventional formalism of field theories would allow us to climb higher and higher up the ladder of energy scales, but we would never reach the top. We would require a theory of a fundamentally different formal type in order to end the ladder of cutoffs. This is presumably the view that Wallace holds, as he claims that if we replace one field theory with another applicable at higher energies, "that field theory in turn will need some kind of short-distance cutoff" (2011, p. 118).

As great as it may be to have a framework in which theories limit their own domain of

applicability, this is certainly not a necessary condition that any good formalism need satisfy. Even if AQFT does not contain cutoffs explicitly, this does not make it at odds with CQFT. Many theories that have been useful in the past do not signal their ultimate demise; on the contrary, most are mathematically well-defined well beyond their domain of applicability. For example, classical theories of fluid dynamics treat fluids as classical continua, and these continua are uniform to arbitrary precision. Classical continuum fluid dynamics is a useful theory, and compatible with classical point mechanics, even though classical point mechanics leads one to believe that the continuum is only an approximation—at some point fluid dynamics must break down. There is nothing within the formalism of fluid mechanics that signals its eventual breakdown; rather, the physical systems we model using classical fluid dynamics, as well as the complementary formalism of classical point particles, give us a physical motivation for the eventual breakdown of the formalism. Deeper theories, such as quantum mechanics, also provide grounds for believing in the limited applicability of both of the complementary classical formalisms. Similarly, we can view AQFT as a complementary picture to the formalism of CQFT. Both formalisms rely on the same general physical principles, though they are implemented in different ways. Though the AQFT formalism does not demarcate its domain of applicability in the form of explicit cutoffs, the necessity of some form of cutoff in CQFT provides reason to believe that the AQFT formalism is only approximately mapping the actual physics. Further, whatever extratheoretical grounds we have for taking cutoffs to be physical—typically in the guise of speculative physics beyond the Standard Model—can inform the scale at which we lose faith in the predictions of *both* the AQFT and CQFT formalisms. When one does not view AQFT and CQFT as rival research programs, the two can work together to provide a deeper

physical understanding of high energy physics, and the role of cutoffs is made clearer.

2.3 Physical significance versus being physical

Are cutoffs really that central? The arguments in the previous section assume that the cutoffs required to generate predictions in CQFT are physical, in the sense that they signal a breakdown of QFT. The fact that perturbative calculations within a particular model diverge when the integrals are unbounded does not entail that field theoretic methodology loses physical significance near these bounds. Undoubtedly we have extratheoretical reasons for supposing that the QFTs making up the Standard Model are not accurate to arbitrary energies—at some point gravity will surely play an important role, to say nothing for possible unknown physics at higher energy scales—but this needn't signify a breakdown of QFTs *in general* beyond a cutoff. Nor is this notion built in to the conceptual apparatus of RG methods, as Wallace claims.[4] It remains entirely possible that a QFT built with more terms in its Lagrangian could describe all relevant physics and be well-defined at all energy scales. In fact, the renormalization group procedure presupposes a theory given in terms of a Lagrangian or Hamiltonian with an arbitrary number of terms. These terms are shown to go to zero in the low energy limit (Wilson and Kogut 1974). We know—using the RG methods to determine the flow of coupling constants—that for non-Abelian gauge theories, interactions become weaker at higher energy scales. Total asymptotic freedom would be one way to eliminate cutoffs at

[4] "Wilsons explanation of the renormalisation procedure relies upon *the failure of the QFT to which it is applied* at very short distances. It is then intriguing to ask how to put on a firm conceptual footing a theory which relies for its mathematical consistency on its own eventual failure". (Wallace 2006, 34, emphasis added) Again, this passage can be read in a way that agrees with the arguments of this section. I am attempting to argue against a naive reading, which takes the failure of *one* QFT (i.e., a single form of interaction, encoded in a particular Lagrangian) to signal the failure of QFT methods in general.

high energies. A successor QFT, such as a grand unified theory or supersymmetry, could therefore unite the strong and electroweak coupling constants, while remaining well-defined to arbitrarily high energies.[5] All that RG methods rely on conceptually is the ability to average out behaviour at high energy scales, and this is compatible with many options for high-energy behaviour. First, our theories could be low-energy approximations that break down at higher energy scales. This could be due to a fundamental granularity or discreteness in the more fundamental theory, or due to the absence of terms in the Lagrangian modelling high energy dynamics. Second, we could have a well-defined high energy dynamics that is unimportant at the energy scales with which we are concerned. In any case, RG methods provide no principled grounds for thinking that cutoffs are "real" in the sense of signifying a breakdown of field theories generally. Unlike the breakdown of classical fluid mechanics—for which we have a more fundamental successor theory (quantum mechanics) providing grounds to reject the continuum as merely an approximation—there is as of yet no (empirically successful) fundamental successor theory for which QFT can be considered a continuum approximation.

One of the major reasons for thinking that cutoffs in QFT mark a regime beyond which the methods of QFT can no longer be applied is the success of RG methods originating from CMP (Wallace 2011, Sec. 1). RG methods were initially developed to investigate long range correlations in materials approaching a phase transition. Long range interactions are those most relevant to global transitions of a material, and so RG

[5]Whether a theory can be made well defined for arbitrarily high energies is a distinct issue from the accuracy of that theory's predictions at high energies. It may turn out that Standard Model QFTs can be extended in a consistent way, but that the high energy predictions turn out to be false. This is the case that is argued in Section 2.2 regarding AQFT.

methods average out the unimportant short range behaviour near a critical point. The apparatus of non-relativistic QFT (i.e., functional integrals using Galilean invariant Lagrangians) is used in CMP as an *approximation* to the discrete atomic (or ionic) physical makeup of bulk systems. Given the the CMP field theories are explicitly constructed as approximations to a known underlying lattice model, we know that the field theoretic methods must break down within CMP. RG flow equations are derived by separating field variables φ into low- and high-momentum components $\varphi = \varphi_{low} + \varphi_{high}$ (where the cutoff from low to high is chosen arbitrarily) and averaging over the high momentum modes. The resulting Lagrangian $\mathcal{L}'(\varphi_{low})$ is then manipulated to fall into the same form as the original Lagrangian $\mathcal{L}(\varphi)$. This process is repeated and generates discrete recursive relations between the rescaled coupling parameters in the $(n+1)$th Lagrangian in terms of the nth one. In the limit where the rescalings are continuous, these become differential equations determining the flow of coupling constants under RG. As the flows are taken to zero frequency—equivalent to the infinite spatial limit—only those parameters relevant to phase transitions will remain in the renormalized Lagrangian. One of the most qualitatively interesting features of successively averaging out short distance (and therefore high energy) degrees of freedom is that, no matter how complicated the initial field dynamics are (encoded as a Lagrangian), only the renormalizable terms will contribute to the low energy dynamics of the theory. This implies that a very broad class of higher energy Lagrangians can "reduce" to the relevant dynamics at lower energy scales.

The success of RG methods in CMP lead to their quick application in QFTs (Wilson 1983)[6], since the relevant formalism is shared between the two disciplines. If we choose

[6]Wilson even forms the QFT/statistical mechanics analogy explicitly, though the source analog in that

to endow the RG methods with similar physical significance in QFT, then we can interpret the high energy cutoffs required as marking the domain at which we expect new physics to occur. The problem is that, because RG flows tell us that our low-energy (effective) QFTs are largely insensitive to the dynamical details at higher energies, they provide little insight our guidance into the high energy physics. Though the path to the successor theory isnt apparent given our current QFTs, the up side is that our best QFTs are protected from the details of our ignorance of high energy dynamics.

Where Wallace might be read to err is in the jump from believing that cutoffs have physical relevance in QFTs to believing that cutoffs *are physical*:

> "This, in essence, is how modern particle physics deals with the renormalization problem: it is taken to presage an ultimate failure of quantum field theory at some short lengthscale, and once the bare existence of that failure is appreciated, the whole of renormalization theory becomes unproblematic, and indeed predictively powerful in its own right" (Wallace 2011, p. 119).[7]

The difference is subtle. Cutoffs can be *physically relevant* in that they signal the breakdown of the *particular* theory or model beyond a certain energy scale, but whether cutoffs themselves *are physical* depends on the precise nature of the breakdown. If the

case is a classical Ising model (Wilson and Kogut 1974). Fraser (2016) has provided an in-depth analysis of the elements of the analogies between QFT and the Ising model, as well as the process of describing RG flow.

[7]Or at least this is a jump he is sometimes guilty of. In other places he is more careful to elaborate on this view, and it appears that he at least appreciates the fact that field theoretic methods may not break down at all (Wallace 2006, pp. 43-4). As mentioned in the introduction, this paper is not a critique of Wallace's view explicitly, but of the misleading way of framing AQFT and CQFT as rivals based on their differing treatments of the arbitrarily small; for this reason I aim to clarify the mistakes in a "naive" reading of Wallace.

breakdown can be remedied by adding new terms in the Lagrangian—effectively changing the particular theory, but retaining the field theoretic framework—then the cutoffs signal new physics, but are not themselves physical. If the breakdown is due to the inapplicability of field theoretic methodology beyond that scale, then the cutoffs are themselves physical.[8] Even if one takes the cutoffs to have physical significance, cutoffs needn't *be physical* in this stronger sense.

One possible reason for thinking that cutoffs are physical is based off of reading too much into the analogy with CMP. We know that field theoretic methods are approximations in bulk matter systems—the atomic theory implies that macroscopic matter is composed of discrete components. The analogy between QFT and CMP is based on the use of the same field theoretic formalism in both disciplines, not on a well-grounded physical similarity.[9] Cutoffs are physical in CMP field theory because field theoretic methods have been introduced as an approximation. Given that discrete quantum mechanics of 10^{23} particles is intractable, we sacrifice (a surprisingly small amount of) precision in order to apply the more soluble methods developed in QFT. But the fact that cutoffs signal the breakdown of field approximations in CMP does not imply that the same is true in QFT. The reasons we treat cutoffs as physical in CMP are absent in QFT; there is no empirically successful theory that claims QFT breaks down due to an underlying discreteness of physics near cutoff scales. Speculative physics may posit some underlying structure for which quantum fields are merely an approximation,

[8]Presumably, the failure of field theoretic methodology in general would require some physical granularity at high energies. This is what I mean by the cutoff being physical and is in direct analogy with the case of non-relativistic QFT in CMP.

[9]Fraser (2016) and Fraser and Koberinski (2016) provide two concrete examples of fruitful formal analogies between QFT and CMP. In the former case, it is the RG flow that is formally analogous, while the latter deals with the formal similarities between spontaneous symmetry breaking within the two theories.

but until any of these theories make successful empirical predictions their significance for interpreting QFTs must be limited.

2.4 Why physical cutoffs are also a problem for CQFT

Even though, as I have argued, there is currently no physically motivated reason for supposing cutoffs to be physical, it may be the case that we find such a reason in the future. Perhaps we will need radically different methods from those of field theory to describe physics beyond the Standard Model. There is no shortage of candidates that claim to radically alter our picture of the world—from 11-dimensional string theory to discrete spacetime to the emergent spacetime of loop quantum gravity. Though experimental support for any of these speculative theories would mean that the axioms of any AQFT must be at best only approximations, this does not mean that CQFT would escape unscathed. Any observed violation of Lorentz invariance would signal bad news for both AQFT and CQFT, and the extent to which we choose to reject or salvage the former, we should do the same for the latter.

Though its importance is not encoded in a set of axioms, Poincaré invariance is of central importance to the physical content of CQFT. In constructing QFTs, one starts by writing down a classical Lagrangian to encode the physical content of the theory. The two major constraints on the form of candidate Lagrangians are renormalizability (dealt with above) and Poincaré invariance. Since the Lagrangian is a scalar, it must remain strictly invariant under the action of the Poincaré group on its component fields. All of the fundamental forces—as described by the Standard Model—are encoded in Lagrangians obeying strict Poincaré invariance. If anything qualifies as physically relevant to CQFT, the Lagrangian certainly does; it is the starting point for building a

QFT, and determines the types of fields, their masses, and the particulars of their interactions. A violation of Poincaré invariance at a more fundamental level—be it in a particular physical process or in the structure of some new spacetime picture—undercuts to the same extent the physical significance of *any and all* theories that depend on Poincaré invariance for their formulation. Thus, despite the lack of rigid and precise axioms demanding Poincaré invariance, the physical content of CQFT stands or falls with AQFT.[10]

Once again, the major difference between AQFT and CQFT lies in the formalism. Though the *physical* content of CQFT is built upon Poincaré invariance[11], the formalism is indifferent to the constraints placed upon the Lagrangian. The success of field theoretic methods in CMP is evidence of the flexibility of the formalism; in CMP the Galilean group is taken as the appropriate symmetry group, given the low energies dealt with. In contrast, the formalisms of various AQFTs are constructed around the axioms. Any theorems that rely on exact Poincaré invariance will only hold in the real world if nature is Poincaré invariant.[12] The greater precision of the formalism in AQFT makes it more rigid in this regard.

If violations of Poincaré invariance are problematic for all variants of QFT, should investigators into the foundations of QFT fret if such violations are experimentally

[10]CQFT *methods* could still be useful, but the theoretical framework of CQFT—as encoded in the Standard Model—depends on Poincaré invariance.

[11]Depending on how one views Poincaré invariance, this may seem odd. The specific transformation properties of scalars, vectors, and tensors under the Poincaré group are undoubtedly formal properties of the particular field representations. However, the physical symmetries represented in this way have a physical basis (e.g., rotation invariance implies that the physical system can be modelled the same way when rotated).

[12]Though it isn't always possible, proofs of the form "If Minkowski spacetime then x" are strengthened and made more robust by also showing "If *approximately* Minkowski spacetime then *approximately* x. Given that our best current theories lead us to believe that spacetime is only locally Minkowski, these are the results for which we can have a high degree of confidence in their robustness.

confirmed? No; the experimental success of QFT implies that the world is at least *approximately* Poincaré invariant, and any evidence revealing the limits of that approximation has no bearing on the theory itself. We have good reason to believe that the QFTs in the Standard Model are not the final story: General Relativity implies that strong gravitational effects distort spacetime, and that our spacetime is only ever Minkowski in small patches where gravity is negligible. Though this approximation seems to hold for experiments at the LHC, if we want a theory that gets spacetime symmetries *exactly* correct, QFTs relying on Poincaré invariance will not do the trick. Rather than abandoning foundations of QFT for being approximate at best, investigation should proceed given that QFTs are highly successful within the energy domain currently testable. To this extent, we are justified in viewing the world as approximately described by QFTs, and should content ourselves with investigating an incomplete (though highly accurate) picture of nature. Whether we are dealing with a formalism that encodes Poincaré invariance into its axiomatic framework, or a formalism in which Poincaré invariance has been used indirectly to construct empirically successful theories, we should not take violations of Poincaré invariance as signalling the failure of either approach. Any robust results obtained within either formalism will still hold approximately, and should be equally subject to foundational analysis.

3 Conclusions

I have tried to show that cutoffs do not provide physical grounds for separating AQFT and CQFT as rival research programs. First, RG methods can be incorporated into AQFT without major issue, and cutoffs can be introduced as well—though explicit

cutoffs provide a more pressing conceptual revision to AQFT. Second, we needn't take AQFT to be an exact description of the world. In the same way that classical fluid dynamics is compatible with classical point mechanics, AQFT defined to arbitrary precision can be compatible with a CQFT that requires cutoffs. The apropriate lesson is that we should take AQFT to be approximately true in sufficiently low energy domains. Finally, even if cutoffs are of physical significance, they don't require a breakdown of continuum methods in general. This idea stems from pushing an analogy with CMP, which appears to be unjustified.

Though the Fraser-Wallace debate has spawned increased investigations into the foundations of QFT, it has set the boundaries of the debate in such a way as to create a false dichotomy: one is forced to choose whether to immerse oneself in the AQFT or CQFT formalisms. When we discard the false dichotomy and recognize AQFT as complementary to CQFT, we open the door to the synthesis of axiomatic methods with Lagrangian QFT. In this way the general features of QFTs can be investigated rigorously in AQFT, and we can be confident that—insofar as the axioms of AQFT capture the physical assumptions of CQFT—the results carry over to CQFT.

Though it is true that there do not yet exist AQFT models that incorporate interactions in four-dimensional spacetime, the successes of AQFT have been compatible with CQFT. Free field theories and ϕ^4_2 interaction theories constructed in AQFT give predictions in agreement with comparable CQFTs. Insofar as AQFT is a successful formalism, its results should be thought of as complementary to those of CQFT: one uses the same physical principles to construct differing formalisms.

In essence, I advocate for a position similar to Wallace's earlier view (though note that in this passage he refers only to specific results of AQFT, such as the spin-statistics

theorem):

> the foundational results which have emerged from AQFT have been of considerable importance in understanding QFT and in general they apply also to Lagrangian QFTs. This paper should be read as complementary to, rather than in competition with, these results (2006, p. 35).

The particular choice of formalism will depend on the scope of the foundational investigation. If the goal is to prove general results applicable to any relativistic QFT, then AQFT is the appropriate formalism; if the goal is to determine the consequences of specific physical interactions, then CQFT should be used.

References

Buchholz, Detlev and Rainer Verch (1995). "Scaling algebras and renormalization group in algebraic quantum field theory". In: *Reviews in Mathematical Physics* 7.8, pp. 1195–1239.

Fraser, Doreen (2009). "Quantum field theory: Underdetermination, inconsistency and idealization". In: *Philosophy of Science* 76, pp. 536–567.

— (2011). "How to take particle physics seriously: A further defence of axiomatic quantum field theory". In: *Studies in History and Philosophy of Modern Physics* 42, pp. 126–135.

— (2016). "The development of renormalization group methods for particle physics: Formal analogies between classical statistical mechanics and quantum field theory". Forthcoming in The British Journal for the Philosophy of Science.

Fraser, Doreen and Adam Koberinski (2016). "The Higgs mechanism and superconductivity: A case study of formal analogies". Forthcoming in Studies in the History and Philosophy of Modern Physics.

Haag, Rudolf and Daniel Kastler (1964). "An algebraic approach to quantum field theory". In: *Journal of Mathematical Physics* 5.7, pp. 848–861.

Halvorson, Hans and Michael Müger (2007). "Algebraic quantum field theory". In: *Handbook of the Philosophy of Physics, Part A*. Ed. by Jeremy Butterfield and John Earman. Elsevier.

Ruetsche, Laura (2011). *Interpreting quantum theories*. Oxford University Press.

Wallace, David (2006). "In defence of naiveté: The conceptual status of Lagrangian quantum field theory". In: *Synthese* 151, pp. 33–80.

Wallace, David (2011). "Taking particle physics seriously: A critique of the algebraic approach to quantum field theory". In: *Studies in History and Philosophy of Modern Physics* 42, pp. 116–125.

Wilson, Kenneth (1983). "The renormalization group and critical phenomena". In: *Reviews of Modern Physics* 55.3, pp. 583–600.

Wilson, Kenneth and John Kogut (1974). "The renormalization group and the ϵ expansion". In: *Physics Reports* 12.2, pp. 77–199.

On Epistemically Detrimental Dissent:

Contingent Enabling Factors v. Stable Difference-Makers.

Soazig Le Bihan and Iheanyi Amadi

Abstract.

The aim of this paper is to critically build on Justin Biddle and Anna Leuschner's characterization (2015) of epistemologically detrimental dissent (EDD) in the context of science. We argue that the presence of non-epistemic agendas and severe non-epistemic consequences are neither necessary nor sufficient conditions for EDD to obtain. We clarify their role by arguing that they are contingent enabling factors, not stable difference-makers, in the production of EDD. We maintain that two stable difference-makers are core to the production of EDD: production of skewed science and effective public dissemination.

Introduction.

The aim of this paper is to critically build on Justin Biddle and Anna Leuschner's characterization of epistemologically detrimental dissent (EDD) in the context of science (2015). We follow their lead in taking 'dissent' to be a particular kind of criticism, i.e. the act of objecting to a widely held conclusion. When done properly, dissent is welcome within scientific practice. As Helen Longino has clearly established, "scientific knowledge is produced collectively through the clashing and

meshing of a variety of points of view (1990, 69). Criticism, when done properly, is integral to the collective advancement of science.[1] Dissent, when an instance of proper criticism, is thus epistemically valuable in the context of science.

Now there are some instances of dissent that come out as epistemically detrimental. That is to say, some instances of dissent seem to impede, not promote, the collective advancement of science. Many examples come to mind, that have been well described in the recent literature (Oreskes and Conway 2010, Biddle and Leushner 2015, Harker 2015). Roughly speaking, EDD is about manufacturing controversy in a particular scientific field. The typical story goes something like the following. The research involved has some severe non-epistemic consequences in terms of, on one side, industry profit, and, on the other side, public welfare; large amounts of money are invested by industry-related groups to (1) produce some skewed research, (2) largely publicize the results through the media, (3) produce an atmosphere of confusion and doubt within the public, (4) launch some campaign against the lead scientists of the field in the media and political world (often through personal attacks and threats); this results in an atmosphere in which the scientists subjectively feel a lot of pressure and discomfort, and also objectively waste precious time and limited resources to address the well-publicized skewed research. At this point, the collective advancement of science is clearly impeded. We have an instance of EDD.

[1] Longino (1990) offers an account of some of the various kinds of epistemically beneficial criticism within science.

The aim of this paper is to properly distinguish, in that story, between (1) contingent enabling factors, and (2) stable difference-makers, in the production of EDD. Our most contentious claim is that the intrusion of non-epistemic agendas and presence of severe non-epistemic risks are contingent enabling factors, not stable difference-makers for EDD. We maintain that two stable difference-makers are core to the production of EDD: production of skewed science and effective public dissemination.

In Section 1, we offer what we take to be the most straightforward argument for the claim that intrusion of non-epistemic agendas is not sufficient in the production of EDD: it may lead to EDD only if it leads to skewed science. In Section 2 we argue that it is not necessary either. Section 3 is devoted to a clarification of the role of intrusion of non-epistemic agendas in EDD on the basis of a distinction between contingent enabling factors and stable difference-makers. Section 4 investigates the consequences of our analysis for the Inductive Risk Account of EDD proposed by Biddle and Leuschner (2015).

Section 1. Non-epistemic agendas: not sufficient for EDD

That intrusion of non-epistemic agendas is not sufficient to the production of EDD has been discussed by Wilholt (2009), and Biddle and Leuschner (2015). Roughly, the point is simply that, unless intrusion of background non-epistemic agendas is such that the work produced *fails to satisfy some of the conventional standards for proper science*, there is no problem. We offer here what we take to be the most straightforward argument for this point.

As the community of philosophers of science have recently come to recognize, intrusion of non-epistemic values in scientific practice is quite common (Douglas 2009). Now obviously, that does not necessarily result in skewed science. If a scientist defends a conclusion C on the basis of evidence E, the fact that some background non-epistemic values enters in her reasoning does not matter if (1) she can publicly produce a reasoning in defense of C, and if (2) that reasoning can be assessed as adequate scientific reasoning by her peers, including peers who do not share the same background non-epistemic values. If these two conditions are met, then the conventional standards for proper science are met, and we do not have a case of skewed science. Now if proper scientific work was produced, there is no a priori reason to think that her work cannot partake in the collective advancement of scientific knowledge. It might do so at various degrees, but that will depend on its heuristic value, which is a priori unrelated to whether or not there was intrusion of non-epistemic values.

Let us push this line of argument a little further. It is important here to underline the fact that the reasoning rendered public by the scientist might not be the actual reasoning through which she came to accept either E or its relevance with regard to C. From a subjective point of view, for example, she might well have had accepted C well before she produced E and the reasoning defending the relevance of E as supporting C. She might well have accepted C for non-epistemic, value-laden, reasons. However, such considerations over the subjective state of scientists do not matter. The collective assessment of scientific research is not in the business of mind reading. No matter what kind of reasoning (or non-reasoning) actually

brought a scientist to believe C, the relevant question is whether she is capable of producing a reasoning in defense of E and its relevance with regard to C that can be publicly, and positively, assessed by the experts in her field. To put it bluntly: the most biased and ill-intentioned scientists are a priori capable of producing good scientific work.[2]

This line of argument applies to the production of dissenting views. Dissenting claims proposed by scientists motivated by non-epistemic agendas do not necessarily lead to skewed science and hence to of EDD. If a reasoning can be publicly produced, and if the members of the scientific community, including members of that community who do not share the same values as the dissenting views' proponents, assess that reasoning as scientifically adequate, then we do not have an instance of skewed dissent. As an instance of work that satisfies the agreed-upon standards of proper scientific practice, the dissenting view could well participate in the advancement of scientific knowledge. It could do so at various degrees, depending on how important the dissenting views are, but that would not depend on whether or not the dissenting views are the product of scientists with non-epistemic agendas. Considerations about the subjective intentions, or background beliefs, of the scientists are irrelevant, unless one can show that skewed science was produced.

[2] This is not denying the actuality of implicit bias. By definition, implicit bias is still bias. As such, it can be recognized by the scientific community for what it is. What is implicit about it is that the biased author (and possibly some of her peers as well) is not even realizing her own bias.

Section 2 Non-epistemic agendas: not necessary for EDD

At this point, we have shown that intrusion of non-epistemic agendas do not necessarily result in the production of EDD. Note that EDD does not require intrusion of non-epistemic agendas either. What would it take to have a case of EDD without any intrusion of non-epistemic agendas? We know that EDD is about manufacturing controversy within a scientific field. First, the controversy is "manufactured", not genuine, because the dissenting view is not based on proper science; it violates some of the commonly accepted standards for proper scientific practice; it is an instance of skewed science. Now skewed science can come to be in many ways. It does not have to result from the intrusion of non-epistemic agendas. One can imagine the case of a scientist, say Jack, who is genuinely interested in partaking in the collective advancement of scientific knowledge, but is also a poor scientist. One can imagine that Jack is very wealthy, and thus has both the time and financial resources to pursue his research, and produce a large amount of work challenging the commonly held views in a given scientific field. Jack, albeit misguided in many ways, could conceivably do all of this with the "purest" goal in mind.

Now one immediately sees that the production of bad science is not enough to produce EDD. Jack's research is likely to be simply ignored by the scientific community. So what would it take to "manufacture" a controversy on the basis of Jack's research? The answer seems rather straightforward: Jack's research needs to be effectively disseminated, so that scientists feel pressured to respond to Jack's

challenges. The standard avenues for dissemination of scientific research, i.e. peer-reviewed publication, however, are not likely to be an option for Jack, since his work is widely recognized by the community as being of poor scientific quality. He must then bypass these avenues, and manage to effectively disseminate his research among the public. Mass media would be a likely option for this. This in turn forces scientists in the field to waste time and resources to address Jack's research. Hence a case of EDD, with the purest epistemic goal at its source.

The case above might seem far-fetched. One objection could be that, unless some non-epistemic values were at stake, it is unlikely that the media and the public would get interested in Jack's research, and Jack would fail to be able to manufacture the controversy. It might be unlikely, but it is surely conceivable. If Jack's public dissemination machinery is effective enough, (mis-) understandings over the state of research in the field of concern could well have serious repercussions on public funding. Jack could well have a very strong network of communication – he could well be the owner of a very large cable and press network. Repeated reporting on public funding of supposedly controversial science could well spur outrage in the public. "Debates" on mass media would ensue. As soon as the scientists would engage in that conversation, Jack's claims would gain in credibility.[3] At the end, Jack's campaign could well be so effective that scientists

[3] This is a point that Hannah Arendt made clear in her insightful analysis of controversy- and doubt-manufacturing in a completely different context, i.e. the (non-)issue of the reality of the Holocaust during WWII (1966/2010).

would indeed be forced to repeatedly address his research to defend their own. So, intrusion of non-epistemic agendas is not necessary to the production of EDD.

Section 3. Stable Difference-Makers v. Contingent Enabling Factor

From the discussion above, we conclude that intrusion of non-epistemic agendas is neither necessary nor sufficient for the production of EDD. Such a conclusion might strike many as unsatisfactory, however. Isn't it the case that intrusion of non-epistemic agendas was an important factor in the production of the common cases of EDD that we have witnessed over the last 50 years? Some may even want to claim that, as a matter of fact, in all of the cases we know of in recent history, no EDD would have occurred if it were not for the intrusion of non-epistemic agendas. This is an important intuition, and arguably, any satisfactory account of EDD ought to make sense of it. Fortunately, we believe there is a way to do so, that is, by appealing to the distinction between contingent enabling factors and stable difference-makers as discussed by Thomson (2003) and Woodward (2010). Thomson (2003) makes the point (contra many theories of causation) that just because 'E would not have happened without C', it does not follow that 'C has caused E'. She argues that the proposition 'E would not have happened without C' only entails that 'C was physically necessary for E'. Consider her example. John built a bridge over the Rapid River. The Rapid River is notoriously wild, and only John, a master-builder, could have done it. From the bridge being built, it ensues that Smith crosses the river. Now John's building the bridge was physically necessary to Smith's crossing the Rapid River, but most would agree that it is misguided to take it

as a cause for it. John's building the bridge, even if "physically necessary" in the whole process, remains largely irrelevant to Smith's crossing the river. It belongs to the background conditions, or environmental conditions, that make Smith's crossing possible, without causing it in any genuine sense of causation. In Thomson's vocabulary, it is only an enabling factor.

Woodward (2010) is interested in analyzing a similar distinction between the core difference-makers and the background conditions. His analysis is useful to flesh out some of the characteristics of enabling factors à la Thomson.[4] One of intuitions Woodward is trying to capture is that some causal relationships are robust, i.e. insensitive to environmental change, while others are contingent on the presence of a specific environment. To do so, he articulates the notions of "stability".[5] A causal relationship, according to Woodward, is stable if and only if it holds over a wide range of background conditions. Some examples might be useful at this point.

[4] Note that we do not claim (and neither does Woodward) to have unveiled the set of necessary and sufficient conditions for factors to qualify as enabling factors by contrast to stable difference-makers. We will only claim that being enabling factors are typically unstable, and hence, that lack of stability serves as a good indicator for a factor to be only enabling, not causing.

[5] Two other notions are articulated in the article. The notion of proportionality serves to address the issue of the proper levels of explanation. The notion of specificity serves to address the issue of coarse v. fine-grain causal influence.

A paradigmatic example of an unstable relation would be the following.[6] "Star" professor P writes a letter of recommendation for Jane, thanks to which Jane gets a job at university U. She would not have gotten the job without it. Jane meets Joe at U, they get married, and have children. Challenged by the difficulties of coupling an academic career with quality parenting, Jane goes into depression. Now consider the following claim: 'P's writing a letter for Jane caused Jane's depression'. Given the story that is given, there is a sense in which P's writing a letter for Jane enabled Jane's suffering from depression, but there is also a strong sense in which it is misguided to take it as a cause for it. The reason is that the relation between P's writing the letter and Jane's suffering from the disease would cease to hold under many small, contingent, changes in the background conditions for the story (Jane and Joe could not have met, they could have decided to not have children, U could have had a very progressive parental leave policy, etc.). The causal relationship between the letter and the depression is thus highly unstable because it holds only in a very specific environment.

Now contrast this with a paradigmatic example of a stable relation. I turn on the heat under my closed pressure cooker (with some water in it). The pressure goes up and the valve shuts down. Clearly, heating up the pressure cooker is a stable cause of the pressure valve to shut down. Many of the most stable causal relations are backed up by what the kind of generalizations that we take to be the laws of physics, or chemistry. These generalizations hold over a wide range of background conditions.

[6] This example is inspired by Woodward (2010) himself inspired by Lewis (1986).

There are obviously various degrees of stability in between these two extreme cases. Stability is not an all or nothing affair. It might also be difficult to figure out which causal relationships are more or less stable. That said, it could also be worth the effort looking into it, because, how stable a factor is could be a measure of how well we can target change by targeting that factor in a given situation. As Woodward explains (2010, 315): "other things being equal, causal relationships that are more stable are likely to be more useful for many purposes associated with manipulation and control than less stable relationships." Applied to our case, if ultimately we hope to be able to alter the manufacturing of controversy and EDD, it could turn out to be very useful to clarify the causal landscape behind EDD by distinguishing between the contingent enabling factors and the more stable difference-makers.

Thomson's and Woodward's analyses are clearly related. Thomson's bridge example is a clear case of a very unstable causal relationship: it holds only under very specific background conditions (The Rapid River could have been gently, Smith could have decided not to cross the bridge, etc.) Some unstable causal relationships as discussed by Woodward are so at least partially because they are relationships of contingent "physical necessity" à la Thomson. So, a causal factor may be highly unstable, despite being 'necessary' to the causal process, if its influence on the process is highly contingent on a specific environment. No matter how "necessary"

in that sense a factor F is, F being unstable points F being a enabling factor, not a stable difference-maker.[7]

The discussion above allows us to bring home two important points. First, it allows us to identify two stable difference-makers for the production of EDD: the production of skewed scientific research and its effective public dissemination. That the combination of these two factors produces an instance EDD holds over a wide range of conditions. What changes in background conditions would make that causal relation to fail? First, one could think of a world in which scientists could ignore even well-advertised skewed science. For example, that could possibly be the case in a world in which production of scientific research would not depend on getting public founding, or in a world in which the public is generally knowledgeable about (the philosophy of) science, and hence, is able to recognize that the well-

[7] Two points of clarification are in order. First, Woodward convincingly argues that the extent to which a cause is stable is related, but not equivalent to, its distal/proximate character vis à vis the effect. Second, Woodward also argues that stability is not dependent on the level of explanation: degrees of stability are not necessarily to how "reductive" the explanation is. So, our distinction between contingent enabling factors and stable difference-makers is not trivial in the sense that the most stable difference-makers would always be the most proximate causes described at the level of fundamental particles.

advertised science is skewed. Arguably, these do not qualify as small changes in the background conditions for scientific practice.[8]

The second point is a clarification of the role played by the intrusion of non-epistemic agendas in the production of EDD. Intrusion of non-epistemic agendas is not a stable difference-maker for the production of EDD. This is because there is a large range of conditions under which intrusion of non-epistemic agendas do not result in EDD. These include the conditions for all the cases in which intrusion of non-epistemic agendas do no result in skewed science. If we take seriously recent work on science and value, intrusion of non-epistemic values is actually the rule, not the exception within the practice of science (Douglas 2009, Intemann 2001, 2015, and references therein). Note that, if our take on Thomson's and Woodward's analyses is correct, then the claim that intrusion of non-epistemic agendas is not a stable difference-maker but only a contingent enabling factor is consistent with the fact that it has been "physically necessary" in many of the well-known instances of EDD. One can consistently say that, while not a stable difference-maker, it has been an important enabling factor for the production of well-publicized skewed science. Intrusion of non-epistemic agendas has been necessary for some groups to develop an *interest in funding* the production and public dissemination of skewed research.

[8] There is also a possibility that some cases of EDD could come out of seemingly proper science "distracting" the public from the most widely held views within the scientific community. We believe that even in these cases, dissenting views do not entail EDD unless there is violation of some conventional standards for proper science. This interesting issue belongs to another paper.

That said it is important to distinguish between factors that are characterized by this kind of 'necessity' (the bridge or letter kind of necessity) and factors that are true stable difference-makers. It is all the more important that, if one of our goals is to alter the production of EDD, then our analysis suggests that intrusion of non-epistemic agendas is not the proper target. Once again, non-epistemic values are the common rule within the practice of science. A more efficient approach in the prevention of EDD would be to understand the various ways skewed science may be produced. This includes the important discussion on the distinction between legitimate and illegitimate use of non-epistemic values in scientific practice (Hicks 2014, Intemann 2015). This in turn includes an investigation of the mechanisms by which intrusion of non-epistemic values does result in skewed science. Implicit bias might one of these mechanisms. Inductive risk bias, as we shall explain in the next section, is another one. Before we turn to this point, let us take stock.

We have clarified the causal landscape for the production of EDD. We have identified two stable difference-makers – production of skewed science and its effective public dissemination; and we have characterized the important role of intrusion of non-epistemic agendas within science as contingent enabling factors for the production and dissemination of skewed research, hence for EDD.

Section 4. Consequences for the Inductive Risk Account of EDD

Biddle and Leuschner have articulated what they call the "inductive risk account" of EDD (2015). According to this account, the following set of conditions are jointly sufficient for the production of EDD (2015, 273):

Dissent from a hypothesis H is epistemically detrimental if each of the following obtains:

(1) The non-epistemic consequences of wrongly rejecting H are likely to be severe

(2) The dissenting research that constitutes the objection violates established conventional standards.

(3) The dissenting research involves intolerance for producer risks at the expense of public risks.

(4) Producer risks and public risks fall largely upon different parties.

Biddle and Leushner admit that these conditions are not necessarily related to the production of EDD (275):

"We are not arguing that, in all possible worlds, research that meets the conditions of the inductive risk account inhibits the progress of science. It is possible, for example, to organize science and to regulate industry in such a way that dissent that meets these conditions is not widely disseminated, does not acquire political authority, and is not used to attack mainstream scientists. But this is not the way in which science and society are currently organized. Dissent that meets the conditions of the inductive risk account is, given current societal arrangements, likely to inhibit knowledge production, particularly because of the success of political, economic, and ideological interests in structuring the dissemination of research."

We think that the framework used in Section 3 can help clarify the causal landscape for the production of EDD offered in the Inductive Risk Account. Our contention is that Biddle and Leuschner, by focusing on inductive risk, have identified a

particular, important, but still contingent, enabling factor, but have failed to clearly distinguish the proper core of stable difference-makers, for the production of EDD. Let us make that point in more details.

The four conditions above can be seen as dividing into three groups. Condition (2) identifies one of the stable difference-makers – production of skewed science. Conditions (1) and (4) together specify some particular enabling conditions for the formation of non-epistemic agendas – the presence of severe and opposing non-epistemic consequences (SONEC). Condition (3) identifies a mechanism by which intrusion of SONEC-related non-epistemic agendas may enable the production of skewed science. In other words, the inductive risk account of EDD identifies an important series of enabling causes leading to one of the two stable difference-makers we have identified in Section 1-3, i.e. production of skewed science. That series of cause is something like this: from the presence of SONEC to biased inductive risk reasoning, and to skewed science. This is an important contribution to the understanding of EDD precisely because it not only identifies some particular enabling factors (the presence of SONEC) for the formation of epistemic agendas, but also a mechanism by which intrusion of SONEC-related non-epistemic agendas may enable the production of skewed science (via inductive risk bias). Now it is also important to clarify the causal landscape and recognize that fulfillment of Condition (2) is the stable difference-maker which fulfillment of Conditions (1), (4), and then (3) enable as a matter of contingent fact. Biddle and Leuschner seem to have missed that useful distinction.

If our analysis in Section 3 is correct, they also have failed to include the second stable difference-maker for EDD, i.e. effective public dissemination. As they admit in the paper (see quote above), the presence of SONEC obviously does not imply that effective public dissemination will ensue. Conversely, as Jack's case shows, effective public dissemination could well be obtained without the presence of SONEC. How (un-)likely this is obviously is an empirical question. No matter how unlikely, however, it is important for our understanding of EDD to mention effective public dissemination as a core stable difference-maker. The inductive risk account fails to do so. Let us underscore, however, that Biddle and Leuschner once again have identified an important mechanism by which presence of SONEC enables effective public dissemination and the manufacturing of controversy: the presence of SONEC not only enables the production of skewed science, but also the establishment of "sophisticated, private-funded network for disseminating [dissenting] results" (2015, 275).

This brings us to our conclusion on the Inductive Risk Account: Biddle and Leuschner have successfully identified an important contingent enabling factor for EDD, i.e. the presence and influence of SONEC. That said, they have failed to distinguish between the different roles that enabling factors and stable difference-makers play in the production of EDD. We hope to have clarified the situation.

Conclusion

Well-known cases of EDD seem to have in common various forms of intrusion of non-epistemic, often SONEC-related, agendas within the science. We have argued

that such intrusion is not core to the production of EDD: neither necessary nor sufficient, it is also not a stable difference-maker. We have clarified its causal role: intrusion of non-epistemic agendas is a contingent enabling factor. Reduced to its core, EDD is just well-advertised bad science. Because it is well advertised, it has an impact on the collective building of scientific knowledge. Because it is bad science, it does not advance that endeavor, but any case negatively impacts it instead.

To make the distinction between contingent enabling factors and stable difference-makers is important for at least three reasons. First, it is important to clarify the causal landscape that leads to the production of EDD, as it simply increases our understanding of EDD. Second, it might suggest more efficient avenues for targeting change. Finally, it is crucial to make room for the intrusion of non-epistemic values within the science without it being epistemologically detrimental. As the community of philosophers of science comes to recognize that such intrusion is the rule rather than the exception, one must leave conceptual room for a distinction between "legitimate" and "illegitimate" role for non-epistemic values within science (Hick 2014, Intemann 2015).

Bibliography

Arendt, Hannah. 1967/2010. "Truth and Politics." In José Medina and David Wood (eds). *Truth. Engagements Across Philosophical Traditions*. Blackwell: 295-314.

Biddle, Justin B. and Anna Leuschner. 2015. "Climate Skepticism and the Manufacture of Doubt: Can Dissent in Science be Epistemically Detrimental?" *European Journal for Philosophy of Science* 5 (3): 261-278.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

Harker, David. 2015. *Creating Scientific Controversies: Uncertainty and Bias in Science and Society*. Cambridge University Press.

Hicks, Daniel J. 2014. "A New Direction for Science and Values." *Synthese* 191 (14): 3271-3295.

Intemann, Kristen. 2015. "Distinguishing between Legitimate and Illegitimate Values in Climate Modeling." *European Journal for Philosophy of Science* 5 (2): 217-232.

———. 2001. "Science and Values: Are Value Judgments always Irrelevant to the Justification of Scientific Claims?" *Philosophy of Science*: S518.

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Lewis, David. 1986. "Postscript c to 'causation': (insensitive causation)" in: *Philosophical papers*, vol 2. Oxford University Press, Oxford: 184–188

Oreskes, Naomi and Erik M. Conway. 2010. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Publishing USA.

Thomson, Judith Jarvis. 2003. "Causation: Omissions." *Philosophy and Phenomenological Research* 66 (1): 81-103.

Wilholt, Torsten. 2009. "Bias and Values in Scientific Research." *Studies in History and Philosophy of Science* Part A 40 (1): 92-101.

Woodward, James. 2010. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biology & Philosophy* 25 (3): 287-318.

PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association -328-

Literal vs. careful interpretations of scientific theories: the vacuum approach to the problem of motion in general relativity

Dennis Lehmkuhl
Einstein Papers Project and HSS Division,
California Institute of Technology
Email: lehmkuhl@caltech.edu

Forthcoming in *Philosophy of Science* (PSA 2016 Supplement)
Version: September 26, 2016

Abstract

The problem of motion in general relativity is about how exactly the gravitational field equations, the Einstein equations, are related to the equations of motion of material bodies subject to gravitational fields. This paper compares two approaches to derive the geodesic motion of (test) matter from the field equations: 'the T approach' and 'the vacuum approach'. The latter approach has been dismissed by philosophers of physics because it apparently represents material bodies by singularities. I shall argue that a careful interpretation of the approach shows that it does not depend on introducing singularities at all, and that it holds at least as much promise as the T approach. I conclude with some general lessons about careful vs. literal interpretations of scientific theories.

Contents

1 Introduction

It is a bit of an irony that one of the most widely embraced definitions of what it means to be a scientific realist is due to the arch-anti-realist Bas van Fraassen. His definition starts by stating that "Science aims to give us, in its theories, a literally true story of what the world is like".[1] And indeed, scientific realists often see themselves as committed to 'taking scientific theories at face value': if the best theories of particle physics say that quarks exist, then we should believe that they exist; if general relativity tells us that gravity is really just an aspect of spacetime structure, then we should believe it; if quantum mechanics tells us that the world is at its core non-deterministic, then we should believe that too.

The problem is that scientific theories, or at least the theories of modern physics, are not that straightforward with us. They may seem so at first, but if you listen to the details of their respective stories, if you take your time to look under the surface, what exactly we should take them to tell us about the world is far from clear. Murray Gell-Mann, the inventor of the concept of quarks, for a long time did not think that quarks should be interpreted as literally existing; neither did Richard Feynman. Albert Einstein passionately resisted the interpretation of general relativity that says that the gravitational force field of Newtonian theory is ontologically reduced to the geometry of spacetime in general relativity. And of course, there is a long-standing battle in foundations of physics about whether quantum mechanics really does tell us that the world is non-deterministic.[2]

In this paper I shall introduce a new case study that provides further evidence for the position that, whether you are a realist or not, the *literal interpretation* of a scientific theory, especially in physics, can be rather misleading. I will argue that what we should aim for is a *careful interpretation*;

[1] Van Fraassen [1980], p.8.

[2] For a discussion of different interpretations of the quark concept see Pickering [1999], for Einstein's opposition to interpreting general relativity as a geometrization of gravity see Lehmkuhl [2014], and for debate on whether quantum mechanics is really indeterministic see e.g. Saunders et al. [2010].

an interpretation of the theory or model or formalism that engages with its details, both with the details of its mathematical structure and with how it is applied to the natural world. Philosophy of science must be willing to look under the hood.

The case study I want to look at is the so-called problem of motion in the general theory of relativity (GR). It asks about the precise relationship between the two sets of equations that are at the very heart of GR. On the one hand there are the Einstein field equations, which give us the dynamics of the gravitational potential (the metric tensor) $g_{\mu\nu}$:

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} = \kappa_E T_{\mu\nu} \quad . \tag{1}$$

On the other hand, we have the geodesic equation that determines which paths through spacetime are geodesics of the connection $\Gamma^{\nu}{}_{\mu\sigma}$ compatible with the metric $g_{\mu\nu}$:

$$\frac{d^2 x_\tau}{ds^2} + \Gamma^{\tau}{}_{\mu\nu} \frac{dx_\mu}{ds} \frac{dx_\nu}{ds} = 0. \tag{2}$$

In GR, material bodies subject only to gravitational fields are supposed to move on the geodesics determined by equation (2).[3] The problem of motion in GR is the question of whether the equations of motion of matter subject to gravitational fields (2) can be derived from the gravitational field equations (1).

Einstein himself, in his first publication on the topic, a paper co-written with Jakob Grommer and published in 1927, compares different classes of attempts to give such a derivation. In particular, Einstein and Grommer distinguish between two classes of attempts at deriving the geodesic motion of matter from the gravitational field equations, which I will term *the T approach* and *the vacuum approach*, respectively. The T approach starts from the realization that the field equations (1) imply the conservation condition, namely that the covariant divergence of the energy-momentum tensor $T_{\mu\nu}$ vanishes:

$$\nabla^\mu T_{\mu\nu} = 0 \quad . \tag{3}$$

[3]It is a big question which systems are actually included under 'material bodies' here. The minimal position is that only test particles are referred to: particles with negligible extension, spin, and self-gravity. However, many actual bodies can be approximated well by test particles in this sense; planets orbiting a star are an example, as we shall see below.

From this, together with certain conditions on the energy-momentum tensor $T_{\mu\nu}$, the T approach derives that material particles move on time-like geodesics. It is this kind of approach to the problem of motion that philosophers have engaged with almost exclusively up to now.[4]

Einstein and Grommer end up dismissing the T approach, and suggest an alternative path to deriving geodesic motion instead. It is a particular version of a *vacuum approach to the problem of motion*. Einstein and Grommer start from the vacuum form of the Einstein field equations,

$$R_{\mu\nu} = 0 \quad , \tag{4}$$

and attempt to derive that the equations (4) imply that material particles move on geodesics.

To the extent that philosophers have engaged with this approach at all, they have quickly dismissed it because it seems to model material bodies by singularities in spacetime; while singularities, by definition, are not even part of spacetime. However, in this paper I shall argue that this dismissal was far too fast, and that indeed the vacuum approach deserves at least as much attention by philosophers as the T approach. The vacuum approach, despite first appearances, engages more closely with some of the most major predictions of GR: both the prediction of the perihelion of Mercury and the prediction of light bending by the Sun utilise the vacuum approach to the derivation of motion of material systems. Indeed, even the prediction of gravitational waves resulting from a binary black hole merger that was recently confirmed rests on the vacuum field equations, for black holes are described by vacuum solutions.[5]

My argument in this paper will proceed in three steps. First, I will argue that the vacuum approach to the problem of motion promises certain advantages that the T approach lacks. Second, I will argue that the problems of the vacuum approach for which it has been dismissed are artefacts of a too literal interpretation of the formalism and its application to the problem at hand. Third, I will argue that a careful interpretation makes the problems disappear; I will argue that the approach does not need to interpret singularities as representing material bodies.

[4]For a comprehensive review of the early history of this approach see Havas [1989] and Kennefick [2005]; for two particularly beautiful exemplars from within this class of proofs see Geroch and Jang [1975] and Ehlers and Geroch [2004], which are investigated by Brown [2007], Malament [2012], and Weatherall [Forthcoming, 2011].

[5]See Abbott et al. [2016] and references therein.

2 A critical comparison of the two research programmes

I said above that the T approach to the problem of motion proceeds via the fact that the Einstein field equations (1) imply the conservation condition (3), which in turn implies the geodesic motion of matter. However, as Malament [2012] pointed out, the conservation condition by itself is not sufficient to prove that the geodesic equation is the equation of motion of material particles. One of the most general proofs from within the T approach, proposed by Geroch and Jang [1975] and further generalised by Ehlers and Geroch [2004], rests not only on the conservation condition (3), but also on the strengthened dominant energy condition, which states:

> Given any timelike covector ξ_μ at any point in M, $T^{\mu\nu}\xi_\mu\xi_\nu \geq 0$ and either $T^{\mu\nu} = \mathbf{0}$ or $T^{\mu\nu}\xi_\mu$ is timelike.

The first clause is effectively the weak energy condition, which states that the mass-energy-momentum density associated with the body in question is always non-negative. The second clause states that every observer will judge the mass-energy-momentum of the body to propagate along time-like curves only.[6]

It would be rather attractive if we did not have to presume that material particles move on time-like curves to then show that these curves are actually time-like geodesics, and if we did not have to presume that matter cannot have non-negative mass-energy. These are weak assumptions about the nature of matter, but they are assumptions.

The vacuum approach to the problem of motion, on the other hand, aims to make *no* assumptions about the nature of matter and its properties at all, and to still derive that matter moves on geodesics. It starts from the question of whether just knowing the exterior gravitational field of a material body, and how this gravitational field interacts with the gravitational field of its surroundings, is enough to derive that the body will move on a geodesic of the metric surrounding it. Arguably, this programme is far more ambitious than the T approach, for it starts with fewer assumptions.[7] And yet, if successful, it would really fit much better the virtues that philosophers have associated

[6] For more on the interpretation of the strengthened dominant energy condition seeWeatherall [2011], Weatherall [Forthcoming] and especially Curiel [Forthcoming].

[7] One might be tempted to argue that despite first appearances the vacuum approach

with the geodesic theorem(s) in the first place: deriving the inertial motion of matter from knowledge of the dynamics of gravitational fields alone.[8]

Einstein was deeply skeptical of the role of the energy-momentum tensor in GR. Throughout the decades, he emphasised that $T_{\mu\nu}$ provides only a 'phenomenological representation of matter'.[9] In Einstein and Grommer [1927], Einstein elaborates that general relativity with an energy-momentum tensor as a source term on the right-hand side of (1) is just not a complete theory: it does not tell us what kind of matter is present, only that it has a certain mass-energy distribution. This perspective on GR was further strengthened by Tupper [1981, 1982, 1983], who showed that knowing the energy-momentum tensor of a material system does not suffice to tell us what kind of matter is present. For example, one and the same mass-energy-momentum distribution $T_{\mu\nu}$ featuring on the right-hand side of the Einstein equations, and solving the Einstein equations for the same metric, can correspond either to an electromagnetic field or a viscous fluid. Knowing the energy-momentum tensor is just not sufficient to know which of these two material systems it is that interacts with the metric field.

Einstein's aim is then to instead start with the vacuum field equations

starts with more demanding assumptions than the T approach. For the vacuum Einstein equations (4) logically imply that the strengthened dominant energy condition (SDEC) holds for the Ricci tensor $R_{\mu\nu}$. The opposite is not true, so that demanding Ricci flatness is clearly a stronger constraint on the Ricci tensor than demanding that it obeys the SDEC. But concluding from this that the vacuum approach starts from stronger assumptions than the T approach would be a mistake. For the T approach assumes i.) the full Einstein field equations (1); and ii.) that the energy-momentum tensor (and thus the Einstein tensor) adheres to the SDEC. The vacuum approach only assumes the vacuum Einstein equations (4), and thus starts with weaker assumptions than the T approach. However, it might well be that depsite *starting* with weaker assumptions than the T approach, a particular manifestation of the vacuum approach might end up with stronger assumptions than a particular manifestation of the T approach. For example, the 1927 Einstein-Grommer vacuum approach, discussed below, involves, among other demands, a so-called equilibrium condition which is supposed to relate solutions to the non-linear field equations to solutions of the linearized field equations in a particular way; no such demand is included in, say, the Geroch-Jang version of the T approach. Thus, further analysis might well show that Einstein and Grommer use stronger assumptions than Geroch and Jang. Einstein himself would likely have been content with that, as long as it allowed him to avoid the introduction of $T_{\mu\nu}$, for reasons discussed below.

[8]Cf. Brown [2007], p. 141 and 163.

[9]See, for example, Einstein [1922], Einstein to Michele Besso, 11 August 1926 (EA-7-361), and Einstein [1936].

(4), treat material particles as singularities in the metric field,[10] and derive that they move on geodesics of a metric $g_{\mu\nu}$ that solves the vacuum field equations (4) in the region through which the particle moves.

To the extent that philosophers have engaged with this approach at all, they have already dismissed it at this point. The main criticism is that the very idea of the approach is flawed: A singularity is not even part of spacetime. How should it be possible to describe its motion in said spacetime?

Both Torretti and Earman essentially answer that this is not possible and that the whole programme is ill-conceived. Earman [1995], p. 12, writes:[11]

> [S]ingularities in the spacetime metric cannot be regarded as taking place at points of the spacetime manifold M. Thus, to speak of singularities in $g_{\mu\nu}$ as geodesics of the spacetime is to speak in oxymorons.

The most detailed discussion of the Einstein-Grommer paper in the philosophical literature is due to Tamir [2012]. After quoting the above statement by Earman, Tamir goes on to write (p.142):

> The proponent of such a "vacuum-cum-singularity" technique is faced with the rather paradoxical challenge of explaining in what sense we can say that a singular curve (ostensibly constituted by the *missing* points in the manifold) is actually a geodesic of the spacetime from which it is absent. Not only is no metric defined at the singularity, but also technically there are not even spacetime points there: the geodesic does not exist.

Tamir then mentions a key ingredient of the Einstein-Grommer approach, namely the distinction between an 'inner metric' and an 'outer metric'.[12] Einstein and Grommer aim to show that the particle characterized by a

[10] In recent years, the adequate definition of a singularity in GR has been a subject of extensive debate, see e.g. Earman [1995] and Curiel [1999]. For Einstein's thoughts on singularities see Earman and Eisenstaedt [1999]; in the context of the Einstein-Grommer paper Einstein clearly thinks of a singularity in the metric field $g_{\mu\nu}$ as a region where the components of the metric tend to infinity.

[11] For similar statements see Torretti [1996], section 5.8.

[12] There is an interesting relationship between Einstein and Grommer's distinction between inner and outer metric (discussed further in section 3) on the one hand and the later distinction between interior and exterior black hole solutions on the other. I do believe that bringing together results and concepts developed in the context of black hole solu-

singular inner metric moves on geodesics of the non-singular outer metric. Tamir states that the "suggested implication" is that we are to compare a second spacetime whose metric is that of the regular outer metric with the singular first spacetime, and identify the regular geodesic of the second spacetime with the singular curve of the first one. He then argues that the thought that the second singularity-free spacetime can teach us anything about the singular original spacetime is "spurious".

My point in the following will be this. Even if this argument were convincing, its premise (the 'suggested implication' that Einstein and Grommer intended to deduce something about a singular spacetime by comparing it to a non-singular spacetime) is not. I shall argue that by looking at the details of the Einstein-Grommer approach we come to a different interpretation of the approach, one that sheds a completely different light on the alleged presence of singularities. We will see that a careful (rather than literal) interpretation of the vacuum approach, and the Einstein-Grommer paper in particular, does not actually depend on introducing singularities at all.

3 The vacuum approach to the problem of motion

3.1 Two ways of looking at Einstein's model of the Sun-Mercury system

In a way, the story of the vacuum approach to the problem of motion starts in 1915, with Einstein's treatment of the orbit of Mercury around the Sun in the context of GR. It is a two-body problem: a small body (Mercury) with a comparatively small mass orbits a large body (the Sun). Einstein seems to postulate (more on the 'seems' below) that the Sun be represented by what would soon be recognized as an approximation to the Schwarzschild metric. He definitely postulates (!) that Mercury moves on a geodesic of said metric.[13] In a way, the problem of motion in GR is about the question of

tions (a special case of vacuum solutions) on the one hand and the vacuum approach to the problem of motion on the other hand is very promising indeed. I will have to postpone a detailed discussion to a later paper; it will include the problem of motion of a binary black hole, the black hole equivalent of the Sun-Mercury two-body system discussed below.

[13]For a careful analysis of Einstein's Mercury paper and how it rests on the Einstein-Besso manuscript see Earman and Janssen [1993], and Janssen's Editorial Note on the

whether this second postulate is really necessary.

If we now look at Einstein's Mercury paper and recall the kind of criticism that was launched against the vacuum approach to the problem of motion, we may find ourselves feeling puzzled. After all, the Schwarzschild metric is a solution to the vacuum field equations, and it has a singularity at its center.[14] If representing material bodies by singular metrics is so problematic, how does it come about that Einstein [1915] successfully predicted the perihelion motion of Mercury? Why is it not problematic to represent the Sun by the singular Schwarzschild metric?

The answer lies in denying the premise of the question. Einstein's treatment of the Sun-Mercury system should *not* be interpreted as involving him representing the Sun by (an approximation of) the Schwarzschild metric. We *know* that the Sun is a material body with non-vanishing mass-energy, and that it does not have a spacetime singularity at its center. What Einstein really does is to convert the two-body problem Sun-Mercury into a one-body problem, where one body (Mercury) is subject to an external gravitational field. It is the exterior gravitational field of the Sun, *not the Sun itself*, that is represented by the Schwarzschild metric. And that is enough to predict the perihelion of Mercury: we don't need to know what the Sun is made of or what happens in its interior; all that matters is the exterior gravitational field that Mercury is subject to.

Thus, worrying about the singularity at the center of the Schwarzschild metric just misses the point: we do not have to interpret the interior part of the Schwarzschild metric literally, at least not in this application.

In the following I shall argue that we should interpret the appearance of singularities in the Einstein-Grommer vacuum approach to the problem of motion in a similar vein.

3.2 The Einstein-Grommer vacuum approach to the problem of motion

The general scheme of the Einstein-Grommer approach proceeds as follows.[15]

Einstein-Besso manuscript in Vol. 4 of the Collected Papers of Albert Einstein (CPAE).

[14]For the history and interpretation of the Schwarzschild metric and its analytic extensions see Eisenstaedt [1989] and Bonnor [1992].

[15]The genesis of the Einstein-Grommer approach has been a bit of a mystery up to now, as pointed out by Kennefick [2005]. However, the work on the 15th volume of Einstein's collected papers has revealed the context and correspondence leading up to that paper,

1. Reformulate the vacuum Einstein equations in terms of a surface integral over a three-dimensional hyper-surface such that we can ask whether gravitational energy-momentum represented by the pseudo-tensor t^{τ}_{α} passes through the surface.[16]

2. Pick a curve that is supposed to represent the path of a material particle.

3. Impose the linear approximation according to which $g_{\mu\nu} = \eta_{\mu\nu} + \gamma_{\mu\nu}$, i.e. assume that, at least close to the curve, the metric deviates from Minkowski spacetime only slightly.

4. Realise that not all solutions to the linearized field equations will correspond to solutions of the non-linear field equations that the linearized field equations approximate. Argue that in the case where an 'equilibrium condition' for the energy-pseudo-tensor of the gravitational field holds, the $\gamma_{\mu\nu}$ of the linearized field equations *will* solve the full non-linear equations reformulated as a surface integral.[17]

5. Now split the $\gamma_{\mu\nu}$ in the immediate neighborhood of the particle into the 'inner metric $\bar{\gamma}_{\mu\nu}$ that the particle itself gives rise to and the 'outer metric $\bar{\bar{\gamma}}_{\mu\nu}$ that is due to other sources (or lack thereof). Observe that the 'outer metric' is entirely regular, even if extended to the point at which the material particle is supposed to be located.

6. Integrate the surface integral that is equivalent to the vacuum field equations 'around' the curve that is supposed to represent the path of a material particle. For the case where the integration surface is a sphere, the equilibrium condition for t^{τ}_{α} simplifies to $\frac{\partial \bar{\bar{\gamma}}_{44}}{\partial x_{\sigma}} = 0$.

and how it fits into Einstein's overall research program. It is a fascinating story; alas, it will have to wait for a separate paper.

[16]There has been a long debate on whether gravitational energy can be adequately represented by a pseudo-tensor; I will not be able to do it justice here. For some details see the introduction to Volume 8 CPAE for the debate between Einstein, Klein, Levi-Civita and Lorentz, for conceptual analysis Hoefer [2000] and especially Trautmann [1962].

[17]This step is very intricate and it would take me a few pages to do it justice. This point of the Einstein-Grommer paper has not been adressed by the literature at all (neither in physics nor in philosophy); I will argue elsewhere that it sheds new light on Einstein's later doubts as to whether the gravitational wave solutions of the linearized equations correspond to gravitational wave solutions in the full non-linear theory.

7. Conclude that the curve that represents the path of a material particle is a geodesic of the outer metric $\bar{\bar{\gamma}}_{\mu\nu}$.[18]

4 Interpreting the Einstein-Grommer approach to the problem of motion

The reader might think that the argument presented in the last section cannot be a faithful representation of the Einstein-Grommer approach; after all, where is the claim that the material particle is represented by a singularity, the reason the approach was dismissed by Earman and Tamir? Indeed, I have omitted that after step 5 of the argument Einstein and Grommer *do* say that one *could* assume that the inner metric $\bar{\gamma}_{\mu\nu}$ is given by what is effectively a three-dimensional counterpart of the Schwarzschild metric: it is spherically symmetric and has a singularity at the center. And yet, *Einstein and Grommer never use this assumption in their argument.* They call the material particle 'the singularity' all the time, but their argument does not depend on assuming *any* particular form for the inner metric, let alone one that is necessarily singular. As a matter of fact, they do not even mention a concrete candidate metric for the outer metric $\bar{\bar{\gamma}}_{\mu\nu}$; all they need is that $\gamma_{\mu\nu}$ is split into the inner metric $\bar{\gamma}_{\mu\nu}$ and the outer metric $\bar{\bar{\gamma}}_{\mu\nu}$ in such a way that $\bar{\bar{\gamma}}_{\mu\nu}$ is non-singular everywhere.

Note that this does not mean that we *know* that the inner metric $\bar{\gamma}_{\mu\nu}$ is non-singular. We don't know anything about the inner metric, for the argument is independent of $\bar{\gamma}_{\mu\nu}$ having any particular form, just like the derivation of Mercury's perihelion was independent of whether there is a singularity at the center of the Schwarzschild metric that represented the exterior field of the Sun.

With regard to the Sun-Mercury system I argued that we should not interpret the Schwarzschild metric as representing the Sun, but as representing its exterior gravitational field. The part of the Sun that is within the event horizon, including the singularity at the center, should not be taken

[18] Einstein and Grommer then go on to generalise this result to the 'non-stationary case', i.e. the case where it is not demanded that the external gravitational field, to which the particle is subject to, does not change in time. They conclude that in this case, too, the particle will move on a geodesic of the outer metric $\bar{\bar{\gamma}}_{\mu\nu}$ that is a solution to the field equations. For the following this generalisation does not make a difference; I will thus refer only to the stationary scenario described above.

as a representation of the *actual* interior of the Sun, but as a *placeholder* or a *blind spot* within the current description of the Sun-Mercury system: a docking station for a theoretical model of the Sun not included in Einstein's Sun-Mercury model.[19]

Likewise, we should interpret the inner metric $\bar{\gamma}_{\mu\nu}$ in the Einstein-Grommer approach as a placeholder for a representation of matter not included in the current theoretical approach. Sure, you *can* set $\bar{\gamma}_{\mu\nu}$ to be a Schwarzschild-like metric with a singularity at the center. But you don't have to do that to make the Einstein-Grommer argument work, and even if you do make that assumption, you should still take this particular inner metric with a singularity at its center as a placeholder for a representation or theory of matter not yet provided.[20]

But now wait a minute. You might have disliked the occurence of singularities as representations of particles, but at least the singularity (in lieu of a non-vanishing energy-momentum tensor) gave you an idea of *where* in spacetime the particle was supposed to be. True, Earman and Tamir rightly pointed out that the singularity is not actually part of spacetime, and so it can hardly serve to localize the particle in spacetime. Still, you might think that we're throwing the baby out with the bath water by not choosing any inner metric. After all, is it not the case then that the curve we have been focusing on is just *any* curve, without any reason to think of this curve as the curve of a material particle?[21]

Again, I think we can counter this criticism by comparing the Einstein-Grommer approach to Einstein's treatment of the Sun-Mercury system in

[19]Note that there are interior extensions of the Schwarzschild metric that model the interior of the Sun by solutions of the non-vacuum field equations (1), for example by an incompressible perfect fluid. See Bonnor [1992], section 5.

[20]If I had given more historical details, I could have, I believe, shown that Einstein himself saw the occurence of a singularity in the inner metric in exactly this way. This exegetical argument would have started with evidence that, from early on, he saw GR as a theory of the pure gravitational field without any constraints on what kinds of matter give rise to the gravitational field. Furthermore, I would have argued that even in the Einstein-Grommer paper he clearly forbids singularities *outside* of material particles (where the theory is supposed to give an adequate and deterministic representation of gravitational fields) but has no problem with them appearing *inside of* material systems, where the theory can provide at best phenomenological placeholders for a future 'proper' theory of matter anyhow. Thus, for Einstein energy-momentum tensors as alleged representatives of material systems were on a par with singularities: both were only placeholders for a proper theory of matter.

[21]I thank Jim Weatherall for putting this question to me.

Einstein [1915]. What Einstein did there was to assume that Mercury would move on *some* geodesic of the exterior gravitational field produced by the Sun. He calculated an approximation to the external gravitational field of a static, spherically symmetric and asymptotically flat body; this gravitational field he saw as represented by the connection components $\Gamma^{\nu}{}_{\mu\sigma}$ of a metric $g_{\mu\nu}$ which deviated only slightly from the flat Minkowski metric. He then inserted these gravitational field components $\Gamma^{\nu}{}_{\mu\sigma}$ into the geodesic equation (2). He showed that this law contained Newton's first law and Newton's second law with a gravitational potential giving rise to a force as a limiting case, and showed how the resulting Keplerian laws for orbits differ in his theory as compared to its Newtonian limit. In the end, he obtained that according to the new theory the perihelion ϵ of *any* geodesic orbit around the Sun is given by

$$\epsilon = 24\pi^3 \frac{a^2}{T^2 c^2 (1 - e^2)} \tag{5}$$

Here a denotes the length of the semimajor axis of the orbit in question, e its eccentricity, c the speed of light, and T the orbital period of the planet in question. Einstein then *takes the astronomically known values for Mercury*, plugs them into equation (5), and thereby predicts that Mercury's perihelion changes by 43" per century.

Note that there is *nothing* in the theoretical description that singles out any particular path as that of Mercury. There is no theoretical representation of Mercury, no model. All that is there is the assumption that Mercury will move on one of the geodesics of the affine connection determined by the spherically symmetric field of the Sun. A general equation that all possible geodesic orbits have to fulfil is derived. And then *external knowledge* is used to single out one of these orbits as that of Mercury. Einstein trusts that the astronomers have measured the orbital period, the semimajor axis and the eccentricity of Mercury correctly. It is this external knowledge, plugged into his theoretical model, which does not in itself contain a representation of Mercury or its path, that produces the prediction.

In many ways, the whole vacuum approach to the problem of motion is about the question as to whether in this kind of scenario we really have to assume the geodesic equation as the equation of motion of matter over and above the gravitational field equations. Indeed, let us look at the Sun-Mercury system within the 1927 Einstein-Grommer approach. The problem of motion, then, is the question whether Einstein really *had to* introduce the

gravitational field equations (to describe the exterior gravitational field of the Sun) *and* the geodesic equation (to describe the path of Mercury subject to this gravitational field) as separate assumptions.[22] Could he have only assumed the gravitational field equations and *derived* that Mercury moves on a geodesic of the exterior field of the Sun? My point is that, just like in Einstein's 1915 treatment, the 1927 Einstein-Grommer approach does not *need to* commit to a theoretical model that allows us to localise Mercury internally. It is fine to ask whether the exterior gravitational field around a given curve 'forces' that curve to be a geodesic. Just like in the 1915 treatment, Einstein and Grommer could then use *external knowledge* about whether that particular curve is actually the curve of a material object, or of Mercury in particular. No inner metric, no singularity to represent the material body, is actually needed.

Let us take a step back though, for there is an important difference between the structure of Einstein's 1915 treatment of Mercury on the one hand and the 1927 Einstein-Grommer approach on the other. In the Mercury case Einstein had assumed (!) that Mercury moves on a geodesic, i.e. a special kind of curve, and model-external knowledge about the period, eccentricity and semimajor axis of Mercury could then be used to determine which of the many geodesics of the Schwarzschild metric corresponded to the path of Mercury. But in the case of the Einstein-Grommer argument, what is in question is whether we can prove that the path of Mercury, say, is a geodesic. Thus, at first sight it looks as if while the 1915 argument only needed external knowledge to determine which geodesic is that of Mercury, appeal to external knowledge in the Einstein-Grommer case would have to determine a.) that this curve is a geodesic and b.) that it is the curve of a material body.

Einstein and Grommer did not aim to derive both a.) and b.). Instead, while Einstein in 1915 used external knowledge at the end of his argument, Einstein and Grommer in 1927 use it at the beginning. They start out by assuming that a given curve is the curve of a material particle, and then ask whether having a regular outer metric (which solves the vacuum field equations) around the curve means that the curve of this material particle,

[22]Interestingly, Einstein did not yet have the final gravitational field equations in the Mercury paper; he found them a week later, in his fourth paper of November 1915. However, the approximation of the Schwarzschild metric that he uses in the Mercury paper is an approximative solution of both the field equations from the Mercury paper, and of the final Einstein field equations.

given the further conditions summarized in section 3.2, *must be* a geodesic. Rather than finishing the argument by appeal to external knowledge (as in Einstein 1915), the Einstein-Grommer argument starts with an appeal to external knowledge, which singles out a particular curve as that of a material body.[23]

Either way, both in Einstein's 1915 treatment and in the Einstein-Grommer approach there is no reason to interpret the singularity (appearing in the Schwarzschild metric or the inner metric, respectively) literally. In both cases, the singularity should be interpreted to signify a placeholder or a blind spot of the theoretical treatment, rather than something that should be interpreted literally, as referring and approximately true. Indeed, both Einstein's 1915 treatment of the Sun-Mercury system and Einstein's and Grommer's treatment of an arbitrary material particle subject to an external gravitational field work just as well if, in the former case, no interior metric (to describe the interior of the Sun) or, in the latter case, no inner metric (to represent the location of the particle on the curve), is ever specified.

5 Conclusion

I started out by saying that whether we are realists or antirealists, we should aim for a careful interpretation, rather than a literal interpretation, of the scientific theory that we want to be realists or anti-realists about. As a case study, I argued that the vacuum approach to the problem of motion in GR, and the Einstein-Grommer approach in particular, is far more sensible and promising if we interpret the singularities *not as representing* material bodies but as *placeholders* for a representation of material bodies that is not included in the model. Indeed, I argued that the approach does not even need the

[23]There is a further disanalogy between Einstein's 1915 derivation of the perihelion of Mercury and the Einstein-Grommer argument of 1927. In the former the choice of (an approximation) the Schwarzschild metric to represent the exterior gravitational field of the Sun does important work in the derivation of Mercury's perihelion. In the Einstein-Grommer approach, no choice of a concrete outer metric is necessary to derive that the curve of the particle which is surrounded by the outer metric must be a geodesic. The reason for this difference is that the Einstein-Grommer approach aims to be more general; it only aims to derive *that* a material body moves on *some* geodesic of the outer metric. However, note that it is not the case that any outer metric is allowed by the approach: the class of outer metrics that the approach can work with is heavily constrained by steps 2 and 3 of the Einstein-Grommer argument (see section 3.2).

introduction of singularities to represent material bodies; their introduction does not do any work in answering the question at hand.[24]

Given that in their paper Einstein and Grommer seem to take the singularities as representing material bodies, one might wonder whether this allegedly more careful interpretation does not fall prey to the criticism that the careful interpreter presumes to understand the theory/formalism in question better than its originators. This might seem at odds with the realist tenet of taking scientists and science 'seriously'. I do indeed think that putting the Einstein-Grommer paper into its proper historical context by analysing Einstein's correspondence leading up to the paper and by relating it to his overarching research project at the time *would* convincingly show that he subscribed to something very much like the 'placeholder interpretation' I defended above. Showing this in detail will have to wait for a much longer paper, and I do not ask the reader to just take my word for it. So let us say, for the sake of the argument, that Einstein and Grommer did indeed intend the singularities as representatives of material objects in a rather straightforward way. I believe that we should not take *their* word for it either. And neither did Einstein. Just a few years after the Einstein-Grommer paper, in his famed 1933 Spencer lectures at the University of Oxford, Einstein told us in his opening words: "If you wish to learn from the theoretical physicist anything about the methods which he uses, I would give you the following advice: Don't listen to his words, examine his achievements."[25]

In philosophy of science, I believe there is no better way of examining a scientist's achievements than by looking for the best possible interpretation

[24]The argument that we should thus not see a realist as comitted to being a realist *about* the singularities appearing in the Einstein-Grommer paper resonates well with selective or posit realism as introduced by Vickers [2013]. The idea there is that we should only be realists with respect to components of a prediction that 'fuel the success' of the prediction, i.e. that are indispensable in the derivation of what is predicted. Using Vickers' distinction the introduction of a singular inner metric in the Einstein-Grommer approach is an idle rather than a working posit. However, note that the call for careful rather than literal interpretations with which I started is independent of / complementary to aiming for identification of the idle posits in a derivation. For *even if* we had found that the introduction of the singular inner metric did do work in the derivation of geodesic motion could we have argued (with less force) that the singularity should be interpreted as a placeholder for a future theory of matter, as a temporary measure within an effective theory, and thus not as something that we should interpret as possesing as much 'reality' or 'referring power' as the regular outer metric governed by the field equations.

[25]See Einstein [1934], and van Dongen [2010] for a detailed analysis of the text.

of his or her theories. To do that, we have to not just listen to the words of the scientist who created or discovered it; we have to see what the theory *does* in practice, how it is *used*; which of its parts really do the work.

Acknowledgments

I would like to thank my colleagues at Caltech and at the Einstein Papers Project for many discussions about the problem of motion and the Einstein-Grommer approach in particular. Thanks are due especially to Diana Kormos-Buchwald, Frederick Eberhardt, and Daniel Kennefick. I would also like to thank audiences at Caltech, Oxford, Irvine, the BSPS 2016 conference in Cardiff, and at the 8th Quadrennial Pittsburgh Fellows conference in Lund, Sweden for many helpful discussions on the topic. I would like to thank especially Sam Fletcher, David Malament and Jim Weatherall for carefully reading earlier versions of this paper, and for the extremely helpful comments they gave me. Finally, I would like to thank Dana Tulodziecki for pointing my attention to the link between posit realism and what I was saying in this paper.

References

Abbott, B., Abbott, R., Abbott, T., Abernathy, M., Acernese, F., Ackley, K., Adams, C., Adams, T., Addesso, P., Adhikari, R. et al. [2016], 'Observation of gravitational waves from a binary black hole merger', *Physical Review Letters* **116**(6), 061102.

Bonnor, W. [1992], 'Physical interpretation of vacuum solutions of einstein's equations. part i. time-independent solutions', *General relativity and Gravitation* **24**(5), 551–574.

Brown, H. R. [2007], *Physical Relativity. Space-time structure from a dynamical perspective*, Oxford University Press, USA.

Curiel, E. [1999], 'The analysis of singular spacetimes', *Philosophy of Science* pp. S119–S145.

Curiel, E. [Forthcoming], A primer on energy conditions, *in* D. Lehmkuhl,

G. Schiemann and E. Scholz, eds, 'Towards a Theory of Spacetime Theories', Einstein Studies, Birkhäuser.

Earman, J. [1995], *Bangs, crunches, whimpers, and shrieks: Singularities and acausalities in relativistic spacetimes*, Oxford University Press, USA.

Earman, J. and Eisenstaedt, J. [1999], 'Einstein and singularities', *Studies in History and Philosophy of Modern Physics* **30**(2), 185–235.

Earman, J. and Janssen, M. [1993], 'Einstein's explanation of the motion of mercury's perihelion', *Einstein Studies* pp. 129–172.

Ehlers, J. and Geroch, R. [2004], 'Equation of motion of small bodies in relativity', *Annals of Phyiscs* **309**, 232–236.

Einstein, A. [1915], 'Erklärung der perihelbewegung des merkur aus der allgemeinen relativitätstheorie', *Königliche Preussische Akademie der Wisenschaften (Berlin)* .

Einstein, A. [1922], *Vier Vorlesungen über Relativitätstheorie gehalten im Mai 1921 an der Universität Princeton*, F. Vieweg. Reprinted as Vol.7, Doc. 71 CPAE; and in various editions as "The Meaning of Relativity" by Princeton University Press.

Einstein, A. [1934], 'On the method of theoretical physics', *Philosophy of science* **1**(2), 163–169.

Einstein, A. [1936], 'Physics and reality', *Journal of the Franklin Institute* **221**, 349–382. Reprinted in A. Einstein (1976) Ideas and Opinions (New York: Dell Publishers), pp. 283-315.

Einstein, A. and Grommer, J. [1927], 'Allgemeine Relativitätstheorie und Bewegungsgesetz', *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften, Physikalisch-Mathematische Klasse* pp. 2–13.

Eisenstaedt, J. [1989], The early interpretation of the schwarzschild solution, *in* D. H. a. J. Stachel, ed., 'Einstein and the History of General Relativity', Birkhäuser, pp. 1–213.

Geroch, R. and Jang, P. [1975], 'Motion of a body in general relativity', *Journal of Mathematical Physics* **16**, 65–67.

Havas, P. [1989], The early history of the "problem of motion" in general relativity, *in* 'Einstein and the History of General Relativity', Vol. 1, pp. 234–276.

Hoefer, C. [2000], 'Energy conservation in gtr', *Studies in History and Philosophy of Modern Physics* **31**.

Kennefick, D. [2005], 'Einstein and the problem of motion: a small clue', *The universe of general relativity* pp. 109–124.

Lehmkuhl, D. [2014], 'Why einstein did not believe that general relativity geometrizes gravity', *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* **46**, 316–326.

Malament, D. B. [2012], A remark about the "geodesic principle" in general relativity, *in* 'Analysis and Interpretation in the Exact Sciences', Springer, pp. 245–252.

Pickering, A. [1999], *Constructing quarks: A sociological history of particle physics*, University of Chicago Press.

Saunders, S., Barrett, J., Kent, A. and Wallace, D. [2010], *Many worlds? Everett, quantum theory, & reality*, OUP Oxford.

Tamir, M. [2012], 'Proving the principle: Taking geodesic dynamics too seriously in Einstein's theory', *Studies In History and Philosophy of Modern Physics* **43**(2), 137–154.

Torretti, R. [1996], *Relativity and geometry*, Dover Publications.

Trautmann, A. [1962], Conservation laws in general relativity, *in* L. Witten, ed., 'Gravitation: An Introduction to Current Research', John Wiley and Sons.

Tupper, B. [1981], 'The equivalence of electromagnetic fields and viscous fluids in general relativity', *Journal of Mathematical Physics* **22**(11), 2666–2673.

Tupper, B. [1982], 'The equivalence of perfect fluid space-times and magnetohydrodynamic space-times in general relativity', *General Relativity and Gravitation* **15**(1).

Tupper, B. [1983], 'The equivalence of perfect fluid space-times and viscous magnetohydrodynamic space-times in general relativity', *General Relativity and Gravitation* **15**(9).

van Dongen, J. [2010], *Einstein's Unification*, Cambridge University Press, Cambridge.

Van Fraassen, B. C. [1980], *The scientific image*, Oxford University Press.

Vickers, P. [2013], 'A confrontation of convergent realism', *Philosophy of Science* **80**(2), 189–211.

Weatherall, J. O. [2011], 'On the status of the geodesic principle in newtonian and relativistic physics', *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics* **42**(4), 276 – 281.
URL: *http://www.sciencedirect.com/science/article/pii/S1355219811000566*

Weatherall, J. O. [Forthcoming], Inertial motion, explanation, and the foundations of classical spacetime theories, *in* D. Lehmkuhl, G. Schiemann and E. Scholz, eds, 'Towards a Theory of Spacetime Theories', Birkhäuser.

PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association -348-

Holism, or the Erosion of Modularity – a Methodological Challenge for Validation

Draft to be presented at PSA 2016

Johannes Lenhard, Bielefeld University

abstract

Modularity is a key concept in building and evaluating complex simulation models. My main claim is that in simulation modeling modularity degenerates for systematic methodological reasons. Consequently, it is hard, if not impossible, to accessing how representational (inner mathematical) structure and dynamical properties of a model are related. The resulting problem for validating models is one of holism.

The argument will proceed by analyzing the techniques of parameterization, tuning, and kludging. They are – to a certain extent – inevitable when building complex simulation models, but corrode modularity. As a result, the common account of validating simulations faces a major problem and testing the dynamical behavior of simulation models becomes all the more important. Finally, I will ask in what circumstances this might be sufficient for model validation.

1. Introduction

For the moment, imagine a scene at a car racing track. The air smells after gasoline. The pilot of the F1 racing car has just steered into his box and is peeling himself out of the straight cockpit. He puts off his helmet, shakes his sweaty hair, and then his eyes make contact to the technical director with a mixture of anger, despair, and helplessness. The engine had not worked as it should, and for a known reason: the software. However, the team had not been successful in attributing the miserable performance to a particular parameter setting. The machine and the software interacted in unforeseen and intricate ways. This explains the exchange of glances between pilot and technical director. The software's internal interactions and interfaces proved to be so complicated that the team had not been able to localize an error or a bug, rather remained

suspicious that some complex interaction of seemingly innocent assumptions or parameter settings was leading to the insufficient performance.

The story happened in fact[1] and it is remarkable since it displays how invasive computational modeling is into areas that smell most analogous. I reported this short piece for another reason, however, namely because the situation is typical for complex computational and simulation models. Validation procedures, while counting on modularity, run against a problem of holism.

Both concepts, modularity and holism, are notions at the fringe of philosophical terminology. Modularity is used in many guises and is not a particularly philosophical notion. It features prominently in the context of complex design, planning, and building – from architecture to software. Modularity stands for first breaking down complicated tasks into small and well-defined sub-tasks and then re-assembling the original global task with a well-defined series of steps. It can be argued that modularity is the key pillar on which various rational treatments of complexity rest – from architecture to software engineering.

Holism is a philosophical term to a somewhat higher degree and is covered in recent compendia. The Stanford Encyclopedia, for instance, includes (sub-)entries on methodological, metaphysical, relational, or meaning holism. Holism generically states that the whole is greater than the sum of its parts, meaning that the parts of a whole are in intimate interconnection, such that they cannot exist independently of the whole, or cannot be understood without reference to the whole. Especially W. V. O. Quine has made the concept popular, not only in philosophy of language, but also in philosophy of science, where one speaks of the so-called Duhem-Quine thesis. This thesis is based on the insight that one cannot test a single hypothesis in isolation, but that any such test depends on "auxiliary" theories or hypotheses, for example how the measurement instruments work. Thus any test addresses a whole ensemble of theories and hypotheses.

Lenhard and Winsberg (2010) have discussed the problem of confirmation holism in the context of validating complex climate models. They argued that "due to interactivity, modularity does not break down a complex system into separately manageable pieces." (2010, 256) In a sense, I want to pick up on this work, but put the thesis into a much more general context, i.e. pointing

[1] In spring 2014, the Red Bull team experienced a crisis due to recalcitrant problems with the Renault engine, due to a partial software update.

out a dilemma that is built on the tension between modularity and holism and that occurs quite generally in simulation modeling. The potential philosophical novelty is debated controversially in philosophy of science, for instance Humphreys (2009) vs. Frigg and Reiss (2009). The latter authors deny novelty, but concede issues of holism might be an exception. My paper confirms that holism is a key concept when reasoning about simulation. (I see more reasons for philosophical novelty, though.)

My main claim is the following: According to the rational picture of design, modularity is a key concept in building and evaluating complex models. In simulation modeling, however, modularity erodes for systematic methodological reasons. Moreover, the very condition for success of simulation undermines the most basic pillar of rational design. Thus the resulting problem for validating models is one of (confirmation) holism.

Section 2 discusses modularity and its central role for the so-called rational picture of design. Herbert Simon's highly influential parable of the watchmakers will feature prominently. It paradigmatically captures complex systems as a sort of large clockwork mechanism. This perspective suggests the computer would enlarge the tractability of complex systems due to its vast capacity for handling (algorithmic) mechanisms. Complex simulations then would appear as the electronic incarnation of a gigantic assembly of cogwheels. This viewpoint is misleading, I will argue. Instead, I want to emphasize the dis-analogy to how simulation models work. The methodology of building complex simulation models thwarts modularity in systematic ways. Simulation is based on an iterative and exploratory mode of modeling that leads to a sort of *holism that erodes modularity*.

I will present two arguments for the erosion claim, one from parameterization and tuning (section 3), the other from klu(d)ging (section 4). Both are, in practice, part-and-parcel of simulation modeling and both make modularity erode. The paper will conclude by drawing lessons about the limits of validation (section 5). Most accounts of validation require, if often not explicitly, modularity and are incompatible with holism. In contrast, the exploratory and iterative mode of modeling restricts validation, at least to a certain extent, to testing (global) predictive virtues. This observation shakes the rational (clockwork) picture of design and of the computer.

2. The rational picture

The design of complex systems has a long tradition in architecture and engineering. At the same time, it has not been much covered in literature, because design was conceived as a matter for experienced craftsmanship rather than analytical investigation. The work of Pahl and Beitz (1984, plus revised editions 1996, 2007) gives a relatively recent account of design in engineering. A second, related source for reasoning about design is the design of complex computer systems. Here, one can find more explicit accounts, since the computer led to complex systems much faster than any tradition of craftsmanship could grow. A widely read example is Herbert Simon's "Sciences of the Artificial" (1969). Still up to today, techniques of high-level languages, object-oriented programming, etc. make the practice of design change on a fast scale.

One original contributor to this discussion is Frederic Brooks, software and computer expert (and former manager at IBM) and also hobby architect. In his 2010 monograph "The Design of Design", he describes the rational model of design that is widely significant, though it is much more often adopted in practice than explicitly formulated in theoretical literature. The rational picture starts with assuming an overview of all options at hand. According to Simon, for instance, the theory of design is the general theory of search through large combinatorial spaces (Simon 1969, 54). The rational model then presupposes a utility function and a design tree, which are spanning the space of possible designs. Brooks rightly points out that these are normally not at hand. Nevertheless, design is conceived as a systematic step-by-step process. Pahl and Beitz aim at detailing these steps in their rational order. Also, Simon presupposes the rational model, arguably motivated by making design feasible for artificial intelligence (see Brooks 2010, 16). Wynston Royce, to give another example, introduced the "waterfall model" for software design (1970). Royce was writing about managing the development of large software systems and the waterfall model consisted in following a hierarchy ("downward"), admitting to iterate steps on one layer, but not with much earlier ("upward") phases of the design process. Although Royce actually saw the waterfall model as a straw man, it was cited positively as paradigm of software development (cf. Brooks on this point).

Some hierarchical order is a key element of the rational picture of design and presumes modularity. Let me illustrate this point. Consider first a simple brick wall. It consists of a multitude of modules, each with certain form and static properties. These are combined into

potentially very large structures. It is a strikingly simple example, because all modules (bricks) are similar.

A more complicated, though closely related, example is the one depicted in figure 1 where an auxiliary building of Bielefeld University is put together from container modules.

Figure 1: A part of Bielefeld University is built from container modules.

These examples illustrate how deeply ingrained modularity is in our way of building (larger) objects. Figure 2 displays a standard picture for designing and developing complex (software) systems.

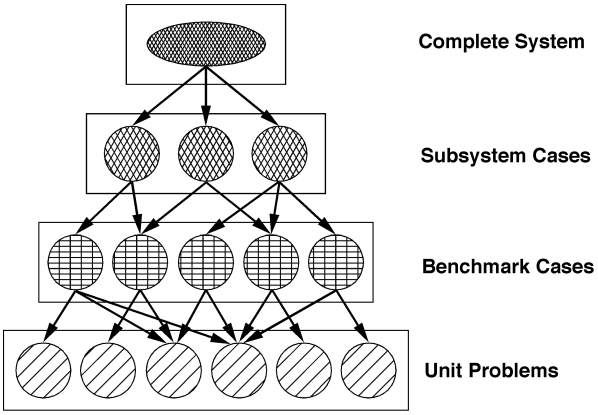


Figure 2: Generic architecture of complex software, from the AIAA Guide for the Verification and Validation of Computational Fluid Dynamics Simulations (1998). Modules of one layer might be used by different modules on a higher layer.

Some complex overall task is split up into modules that can be tackled independently and by different teams. The hierarchical structure shall ensure the modules can be integrated to make up the original complex system. Modularity not only plays a key role when designing and building complex systems, it also is of crucial importance when taking account of the system. Validation is usually conceived in the very same modular structure: independently validated modules are put together in a controlled way for making up a validated bigger system. The standard account of how computational models are verified and validated gives very rigorous guidelines that are all based on the systematic realization of modularity (Oberkampf and Roy 2010, see also Fillion 2017). In short, modularity is key for designing as well as for validating complex systems.

This observation is paradigmatically expressed in Simon's parable of the two watchmakers. You find it in Simon's 1962 paper "The Architecture of Complexity" that has become a chapter in his immensely influential "The Sciences of the Artificial" (Simon 1969). There, Simon investigates the structure of complex systems. The stable structures, so Simon argues, are the hierarchical ones. He expressed his idea by telling the parable of the two watchmakers named Hora and Tempus (1969, 90-92). P. Agre describes the setting with the following words:

"According to this story, both watchmakers were equally skilled, but only one of them, Hora, prospered. The difference between them lay in the design of their watches. Each design involved 1000 elementary components, but the similarity ended there. Tempus' watches were not hierarchical; they were assembled one component at a time. Hora's watches, by contrast, were organized into hierarchical subassemblies whose "span" was ten. He would combine ten elementary components into small subassemblies, and then he would combine ten subassemblies into larger subassemblies, and these in turn could be combined to make a complete watch." (Agre 2003)

Since Hora takes additional steps for building modules, Tempus' watches need less time for assembly. However, it was Tempus' business that did not thrive, because of an additional condition not yet mentioned, namely some kind of noise. From time to time the telephone rings and whenever one of the watchmakers answers the call, all cogwheels and little screws fall apart and he has to re-start the assembly. While Tempus had to start from scratch, Hora could keep all finished modules and work from there. In the presence of noise, so the lesson goes, the modular

strategy is by far superior. Agre summarizes that modularity, he speaks of the functional role of components, comes out as a necessary element when designing complex systems:

"For working engineers, hierarchy is not mainly a guarantee that subassemblies will remain intact when the phone rings. Rather, hierarchy simplifies the process of design cognitively by allowing the functional role of subassemblies to be articulated in a meaningful way in terms of their contribution to the function of the whole. Hierarchy allows subassemblies to be modified somewhat independently of one another, and it enables them to be assembled into new and potentially unexpected configurations when the need arises. A system whose overall functioning cannot be predicted from the functionality of its components is not generally considered to be well-engineered." (Agre 2003)

Now, the story works with rather particular examples insofar as watches exemplify complicated mechanical devices. The universe as a giant clockwork has been a common metaphor since the seventeenth century. Presumably, Simon was aware the clockwork picture is limited and he even mentioned that complicated interactions could lead to a sort of pragmatic holism.[2] Nonetheless, the hierarchical order is established by the interaction of self-contained modules.

There is an obvious limit to the watchmaker picture, namely systems have to remain manageable by human beings (watchmakers). There are many systems of practical interest that are too complex – from the earth's climate to the aerodynamics of an airfoil. Computer models open up a new path here, since simulation models might contain a wealth of algorithmic steps far beyond what can be conceived in a clockwork picture.[3] From this point of view, the computer appears as a kind of amplifier that helps to revitalize the rational picture. Do we have to look at simulation models as a sort of gigantic clockworks? In the following, I will argue that this viewpoint is seriously misleading. Simulation models are different from watches in important ways and I

[2] This kind of holism hence can occur even when modules are "independently validated", since these modules when connected together could interact with each other in unpredicted ways. This is a strictly weaker form of holism than the one I am going to discuss.

[3] Charles Babbage had designed his famous „Analytical Engine" as a *mechanistic* computer. Tellingly, it did encounter serious problems exactly because of the mechanical limitations of its construction.

want to focus on the dis-analogy.[4] Finally, we will learn from the investigation of simulation models about our picture of rationality.

3. Erosion of modularity 1: Parameterization and tuning

In stark contrast to the cogwheel picture of the computer, the methodology of simulation modeling erodes modularity in systematic ways. I want to discuss two separate though related aspects, firstly, parameterization and tuning and, secondly, kluging (also called kludging). Both are, for different reasons, part-and-parcel of simulation modeling; and both make modularity of models erode. Let us investigate them in turn and develop two arguments for erosion.

Parameterization and tuning are key elements of simulation modeling that stretch the realm of tractable subject matter much beyond what is covered by theory. Furthermore, simulation models can make predictions even in fields that *are* covered by well-accepted theory only with the help of parameterization and tuning. In this sense, the latter are success conditions for simulations.

Before we start with discussing an example, let me add a few words about terminology. There are different expressions that specify what is done with parameters. The four most common ones are (in alphabetical order): adaptation, adjustment, calibration, and tuning. These notions describe very similar activities, but also valuate differently what parameters are good for. Calibration is commonly used in the context of preparing an instrument, like calibrating a scale one time for using it very often in a reliable way. Tuning has a more pejorative tone, like achieving a fit with artificial measures, or fitting to a particular case. Adaptation and adjustment have more neutral meanings.

Atmospheric circulation is a typical example. It is modeled on the basis of accepted theory (fluid dynamics, thermodynamics, motion) on a grand scale. Climate scientists call this the "dynamical core" of their models and there is more or less consensus about this part. Although the employed theory is part of physics, climate scientists mean a different part of their models when they speak of "the physics". It includes all the processes that are not completely specified from the dynamical core. These processes include convection schemes, cloud dynamics, and many more.

[4] There are several dis-analogies. One I am not discussing is that clockworks lack multi-functionality.

The "physics" is where different models differ and the physics is what modeling centers regard as their achievements and try to maintain even if their models change into the next generation.

The physics acts like a specifying supplement to the grand scale dynamics. It is based on modeling assumptions, say which sub-processes are important in convection, what should be resolved in the model and what should be treated via a parameterization scheme. Often, such processes are not known in full detail, and some aspects (at least) depend on what happens on a sub-grid scale. The dynamics of clouds, for instance, depends on a staggering span of very small (molecular) scales and much larger scales of many kilometers. Hence even if the laws that guide these processes would be known, they could not be treated explicitly in the simulation model. Modeling the physics has to bring in parameterization schemes.[5]

How does moisture transport, for example, work? Rather than trying to investigate into the molecular details of how water vapor is entrained into air, scientists use a parameter, or a scheme of parameters, that controls moisture uptake so that known observations are met. Often, such parameters do not have a direct physical interpretation, nor do they need one, like when a parameter stands for a mixture of processes not resolved in the model. The important property rather is that they make the parameterization scheme flexible, so that the parameters of such a scheme can be changed in a way that makes the properties of the scheme (in terms of climate dynamics) match some known data or reference points.

From this rather straightforward observation follows an important fact. A parameterization, including assignments of parameter values, makes sense only in the context of the larger model. Observational data are not compared to the parameterization in isolation. The Fourth Assessment Report of the IPCC acknowledges the point that "parameterizations have to be understood in the context of their host models" (Solomon et al. 2007, 8.2.1.3)

The question of whether the parameter value that controls moisture uptake (in our oversimplified example) is adequate can be answered only by examining how the entire parameterization behaves and, moreover, how the parameter value in the parameterization in the larger simulation model behaves. Answering such questions would require, for instance, looking at more global properties like mean cloud cover in tropical regions, or the amount of rain in some area. Briefly

[5] Parameterization schemes and their more or less autonomous status are discussed in the literature, cf. Parker 2013, Smith 2002, or Gramelsberger and Feichter 2011.

stated, parameterization is a key component of climate modeling and tuning is part-and-parcel of parameterization.[6]

It is important to note that tuning one parameter takes the values of other parameters as given, be they parameters from the same scheme, or be they parts of other schemes that are part of the model. A particular parameter value (controlling moisture uptake) is judged according to the results it yields for the overall behavior (like cloud cover). In other words, tuning is a local activity that is oriented at global behavior. Researchers might try to optimize parameter values simultaneously, but for reasons of computational complexity, this is possible only with a rather small subset of all parameters. A related issue is statistical regression methods that might be caught up in a local optimum. In climate modeling, skill and experience remain to be important for tuning (or adjustment).

Furthermore, tuning parameters is not only oriented at the global model performance, it tends to blur the local behavior. This is because every model will be importantly imperfect, since it contains technical errors, works with insufficient knowledge, etc. – which is just the normal case in scientific practice. Now, tuning a parameter according to the overall behavior of the model then means that the errors, gaps, and bugs get compensated against each other (if in an opaque way). Mauritsen et al. (2012) have pointed this out in their pioneering paper about tuning in climate modeling.

In climate models, cloud parameterizations play an important role, because they influence key statistics of the climate and, at the same time, cover major (remaining) uncertainties about how an adequate model should look like. Typically, such a parameterization scheme includes more than two dozens of parameters; most of them do not carry a clear physical interpretation. The simulation then is based on the balance of these parameters in the context of the overall model (including other parameterizations). Over the process of adjusting the parameters, these schemes become inevitably convoluted. I leave aside that models of atmosphere and oceans get coupled, which arguably aggravates the problem.

[6] The studies of so-called perturbed physics ensembles convincingly showed that crucial properties of the simulation models hinge on exactly how parameter values are assigned (Stainforth et al. 2007).

Tuning is inevitable, part-and-parcel of simulation modeling methodology. It poses great challenges, like finding a good parameterization scheme for cloud dynamics, which is a recent area of intense research in meteorology. But when is a parameterization scheme a good one? On the one side, a scheme is sound when it is theoretically well motivated, on the other side, the key property of a parameterization scheme is its adaptability. Both criteria do not point into the same direction. There is, therefore, no optimum; finding a balance is still considered an art. I suspect that the widespread reluctance against publishing about practices of adjusting parameters comes from reservations against aspects that call for experience and art rather than theory and rigorous methodology.

I want to maintain that nothing in the above argumentation is particular to climate. Climate modeling is just one example out of many. The point holds for simulation modeling quite generally. Admittedly, climate might be a somewhat peculiar case, because it is placed in a political context where some discussions seem to require that only ingredients of proven physical justification and realistic interpretation are admitted. Arguably, this expectation might motivate using the pejorative term of tuning. This reservation, however, ignores the very methodology of simulation modeling. Adjusting parameters is by no means particular to climate modeling, nor is it confined to areas where knowledge is weak.

Another example will document this. Adjusting parameters is also occurring thermodynamics, an area of physics with very high theoretical reputation. The ideal gas equation is even taught in schools, it is a so-called equation of state (EoS) that describes how pressure and temperature depend on each other. However, actually using thermodynamics requires to work with less idealized equations of state than the ideal gas equation. More complicated equations of state find wide applications also in chemical engineering. They are typically very specific for certain substances and require extensive adjustment of parameters as Hasse and Lenhard (2017) describe and analyze. Clearly, being able to process specific adjustment strategies that are based on parameterization schemes is a crucial success condition. Simulation methods have made applicable thermodynamics in many areas of practical relevance, exactly because equations of state are tailored to particular cases of interest via adjusting parameters.

One further example is from quantum chemistry, namely the so-called density functional theory (DFT), a theory developed in the 1960s that won the Nobel prize in 1998. Density functionals

capture the information of the Schroedinger equation, but are much more computationally tractable. However, only many-parameter functionals brought success in chemistry. The more tractable functionals with few parameters worked only in simpler cases of crystallography, but were unable to yield predictions accurate enough to be of chemical interest. Arguably, being able to include and adjust more parameters has been the crucial condition that had to be satisfied before DFT could gain traction in computational quantum chemistry, which happened around 1990. This traction, however, is truly impressive. DFT is by now the most widely used theory in scientific practice, see Lenhard (2014) for a more detailed account of DFT and the development of computational chemistry.

Whereas the adjustment of parameters – to use the more neutral terminology – is pivotal for matching given data, i.e. for predictive success, this very success condition also entails a serious disadvantage.[7] Complicated schemes of adjusted parameters might block theoretical progress. In our climate case, any new cloud parameterization that intends to work with a more thorough theoretical understanding has to be developed for many years and then has to compete with a well-tuned forerunner. Again, this kind of problem is more general. In quantum chemistry, many-parameter adaptations of density functionals have brought great predictive success but at the same time render the rational re-construction of why such success occurs hard, if not impossible (Perdew et al. 2005, discussed in Lenhard 2014). The situation in thermodynamics is similar, cf. Hasse and Lenhard (2017).

Let us take stock regarding the first argument for the erosion of modularity. Tuning, or adjusting, parameters is not merely an *ad hoc* procedure to smoothen a model, rather it is a pivotal component for simulation modeling. Tuning convolutes heterogeneous parts that do not have a common theoretical basis. Tuning proceeds holistically, on basis of global model behavior. How particular parts function often remains opaque. By interweaving local and global considerations, and by convoluting the interdependence of various parameter choices, tuning destructs modularity.

Looking back to Simon's clockmaker story, we see that its basic setting does not match the situation in a fundamental way. The perfect cogwheel picture is misleading, because it presupposes a clear identification of mechanisms and their interactions. In our examples, we saw

[7] There are other dangers, like over-fitting, that I leave aside.

that building a simulation model, different from building a clockwork, cannot proceed top-down. Moreover, different modules and their interfaces get convoluted during the processes of mutual adaptation.

4. Erosion of modularity 2: kluging

The second argument for the erosion of modularity approaches the matter from a different angle, namely from a certain practice in developing software known as kluging (also spelled kludging)[8]. "Kluge" is a term from colloquial language that became a term in computer slang. I remember when back in my childhood our family and another, befriended one drove towards holidays in two cars. In the middle of the night, while crossing the Alps, the exhaust pipe of our friends before us broke, creating a shower of sparks where the pipe met the asphalt. There was no chance of getting the exhaust pipe repaired, but the father did not hesitate long and used his necktie to fix it provisionally.

The necktie worked as a kluge, which is in the words of Wikipedia "a workaround or quick-and-dirty solution that is clumsy, inelegant, difficult to extend and hard to maintain, yet an effective and quick solution to a problem." The notion has been incorporated and become popular in the language of software programming and is closely related to the notion of bricolage.

Andy Clark, for instance, stresses the important role played by kluges in complex modular computer modeling. For him, a kluge is "an inelegant, 'botched together' piece of program; something functional but somehow messy and unsatisfying", it is—Clark refers to Sloman—"a piece of program or machinery which works up to a point but is very complex, unprincipled in its design, ill-understood, hard to prove complete or sound and therefore having unknown limitations, and hard to maintain or extend". (Clark 1987, 278)

Kluges carried forward their way from programmers' colloquial language into the body of philosophy guided by scholars like Clark and Wimsatt who are inspired both by computer

[8] Both spellings „kluge" and „kludge" are used. There is not even agreement of how to pronounce the word. In a way, that fits to the very concept. I will use "kluge", but will not change the habits of other authors cited with "kludge".

modeling and evolutionary theory.[9] The important point in our present context is that kluges may function for a whole system, i.e. for the performance of the entire simulation model, whereas it has no meaning in relation to the submodels and modules: "what is a kludge considered as an item designed to fulfill a certain role in a large system, may be no kludge at all when viewed as an item designed to fulfill a somewhat different role in a smaller system." (Clark 1987, 279)

Since kluging stems from colloquial language and is not seen as a good practice anyway, examples cannot be found easily in published scientific literature. This observation notwithstanding, kluging is a widely occurring phenomenon. Let me give an example that I know from visiting an engineering laboratory. There, researchers (chemical process engineers) are working with simulation models of an absorption column, the large steel structures in which reactions take place under controlled conditions. The scientific details do not matter here, since the point is that the engineers build their model on the basis of a couple of already existing modules, including proprietary software that they integrate into their simulation without having access to the code. Moreover, it is common knowledge in the community that this (unknown) code is of poor quality. Because of programming errors and because of ill-maintained interfaces, using this software package requires modifications on the part of the remaining code outside the package. These modifications are there for no good theoretical reason, albeit for good practical reasons. They make the overall simulation run as expected (in known cases); and they allow working with existing software. The modifications thus are typical kluges.

Again, kluging occurs in virtually every site where large software programs are built. Simulation models hence are a prime instance, especially when the modeling steps of one group build on the results (models, software packages) of other groups. One common phenomenon is the increasing importance of "exception handling", i.e. of finding effective repairs when the software, or the model, performs in unanticipated and undesired ways. In this situation, the software might include a bug that is invisible (does not affect results) most of the time, but becomes effective under particular conditions. Often extensive testing is needed for finding out about unwanted behavior that occurs in rare and particular situations that are conceived of as "exceptions", indicating that researchers do not aim at a major reconstruction, but at a local repair,

[9] The cluster of notions like bricolage and kluging common in software programming and biological evolution would demand a separate investigation. See, as a teaser, Francois Jacob's account of evolution as bricolage (1994).

counteracting this particular exception. Exception handling can be part of a sound design process, but increased use of exception handling is symptomatic of excessive kluging.

Presumably all readers who ever contributed to a large software program know about experiences of this kind. It is commonly accepted that the more comprehensive a piece of software gets, the more energy for exception handling new releases will require. Operating systems of computers, for example, often receive weekly patches. Many scientists who work with simulations are in a similar situation, though not obviously so.

If, for instance, meteorologists want to work on, say, hurricanes, they will likely take a meso-scale (multi-purpose) atmospheric model from the shelf of some trusted modeling center and add specifications and parameterizations relevant for hurricanes. Typically, they will not know in exactly what respects the model had been tuned, and also lack much other knowledge about strengths and weaknesses of this particular model. Consequently, when preparing their hurricane modules, they will add measures into their new modules that somehow balance out undesired model behavior. These measures can also be conceived as kluges.

Why should we see these examples as typical instances and not as exceptions? Because they arise from practical circumstances of developing software, which is a core part of simulation modeling. Software engineering is a field that was envisioned as the "professional" answer to the increasing complexity of software. And I frankly admit that there are well-articulated concepts that would in principle ensure software is clearly written, aptly modularized, well maintained, and superbly documented. However, the problem is that science *in principle* is different from science *in practice*.

In practice, there are strong and constant forces that drive software development into resorting to kluges. Economic considerations are always a reason, be it on the personal scale of research time, be it on the grand scale of assigning teams of developers to certain tasks. Usually, software is developed "on the move", i.e. those who write it have to keep up with changing requirements and a narrow timeline, in science as well as industry. Of course, in the ideal case the implementation is tightly modularized. A virtue of modularity is that it is much quicker incorporating "foreign" modules than developing them from scratch.

If these modules have some deficiencies, however, the developers will usually not start a fundamental analysis of how unexpected deviations occurred, but rather spend their energy for

adapting the interfaces so that the joint model will work as anticipated in the given circumstances. In common language: repair, rather than replace. Examples reach from integrating a module of atmospheric chemistry into an existing general circulation model up to implementing the new version of the operating system of your computer. Working with complex computational and simulation models seems to require a certain division of labor and this division, in turn, thrives on software traveling easily. At the same time, this will provoke kluges on the side of those that try to connect software modules.

Kluges thus arise from unprincipled reasons: throw-away code, made for the moment, is not replaced later, but becomes forgotten, buried in more code, and eventually fixed. This will lead to a cascade of kluges. Once there, they prompt more kluges, tending to become layered and entrenched.[10]

Foote and Yoder, prominent leaders in the field of software development, give an ironic and funny account of how attempts to maintain a rationally designed software architecture constantly fail in practice.

"While much attention has been focused on high-level software architectural patterns, what is, in effect, the de-facto standard software architecture is seldom discussed. This paper examines this most frequently deployed of software architectures: the BIG BALL OF MUD. A big ball of mud is a casually, even haphazardly, structured system. Its organization, if one can call it that, is dictated more by expediency than design. Yet, its enduring popularity cannot merely be indicative of a general disregard for architecture. (…) 2. Reason for degeneration: ongoing evolutionary pressure, piecemeal growth: Even systems with well-defined architectures are prone to structural erosion. The relentless onslaught of changing requirements that any successful system attracts can gradually undermine its structure. Systems that were once tidy become overgrown as piecemeal growth gradually allows elements of the system to sprawl in an uncontrolled fashion." (Foote and Yoder 1999, ch. 29)

I would like to repeat the statement from above that there is no necessity in the corruption of modularity and rational architecture. Again, this is a question of science in practice vs. science in principle. "A sustained commitment to refactoring can keep a system from subsiding into a big

[10] Wimsatt (2007) writes about "generative entrenchment" when speaking about the analogy between software development and biological evolution, see also Lenhard and Winsberg (2010).

ball of mud," Foote and Yoder concede. There are even directions in software engineering that try to counteract the degradation into Foote's and Yoder's big ball of mud. The movement of "clean code", for instance, is directed against what Foote and Yoder describe. Robert Martin, the pioneer of this school, proposes to keep code clean in the sense of not letting the first kluge slip in. And surely there is no principled reason why one should not be able to avoid this. However, even Martin accepts the diagnosis of current practice.

Similarly, Richard Gabriel (1996), another guru of software engineering, makes the analogy to housing architecture and Alexander's concept of "habitability", which intends to integrate modularity and piecemeal growth into one "organic order". Anyway, when describing the starting point, he more or less duplicates what we heard above from Foote and Yoder.

Finally, I want to point out that the matter of kluging is related to what is discussed in philosophy of science under the heading of opacity (like in Humphreys 2009). Highly kluged software becomes opaque. One can hardly disentangle the various reasons that led to particular pieces of code, because kluges are sensible only in the particular context at the time. In this important sense, simulation models are historical objects. They carry around – and depend on – their history of modifications. There are interesting analogies with biological evolution that have become a topic when complex systems had become a major issue in discussion computer use. Winograd and Flores, for instance, come to a conclusion that also holds in our context here: "each detail may be the result of an evolved compromise between many conflicting demands. At times, the only explanation for the system's current form may be the appeal to this history of modification." (1991, 94)[11]

Thus, the brief look into the somewhat elusive field of software development has shown us that two conditions foster kluging. First, the exchange of software parts that is more or less motivated by flexibility and economic requirements. This thrives on networked infrastructure. Second, iterations and modifications are easy and cheap. Due to the unprincipled nature of kluges, their construction requires repeated testing whether they actually work in the factual circumstances. Kluges hence fit to the exploratory and iterative mode of modeling that characterizes

[11] Interestingly, Jacob (1994) gives a very similar account of biological evolution when he writes that simpler objects are more dependent on (physical) constraints than on history, while history plays the greater part when complexity increases.

simulations. Furthermore, layered kluges solidify themselves. They make code hard or impossible to understand; modifying pieces that are individually hard to understand will normally lead to a new layer of kluges – and so on. Thus, kluging makes modularity erode and this is the second argument why simulation modeling systematically undermines modularity.

5. The limits of validation

What does the erosion of modularity mean for the validation of computer simulations? We have seen that the power and scope of simulation is built on the tendency toward holism. But holism and the erosion of modularity are two sides of the same coin. The key point regarding methodology is that holism is driven by the very procedure that makes simulation so widely applicable! It is through adjustable parameters that simulation models can be applied to systems beyond the control of theory (alone). It is through this very strategy that modularity erodes.

One ramification of utmost importance is about the concept of validation. In the context of simulation models the community speaks of verification and validation, or "V&V". Both are related, but the unanimous advice in the literature is to keep them separate. While verification checks the model internally, i.e. whether the software indeed captures what it is supposed to, validation checks whether the model adequately represents the target system. A standard definition states that "verification [is] the process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model." While validation is defined as "the process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model." (Oberkampf and Trucano 2000, 3) Though there is some leeway of defining V&V, you get the gist of it from the saying: verification checks whether the model is right[12], while validation checks whether we have the right model.

Due to the increasing usage and growing complexity of simulations, the issue of V&V is itself a growing field in simulation literature. One example is the voluminous monograph by Oberkampf and Roy (2010) that meticulously defines and discusses the various steps to be included in V&V procedures. A first move in this analysis is to separate model form from model parameters. Each

[12] This sloppy saying should not obscure that the process of verification comprises a package of demanding tasks.

parameter then belongs to a particular type of parameter that determines which specific steps in V&V are required. Oberkampf gives the following list of model parameter types:

> " - measurable properties of the system or the surroundings,
> - physical modeling parameters,
> - ad hoc model parameters,
> - numerical algorithm parameters,
> - decision parameters,
> - uncertainty modeling parameters." (Oberkampf and Roy 2010, section 13.5.1, p.623)

My point is that the adjustable parameters we discussed are of a type that is evading the V&V fencing. These parameters cannot be kept separate from the model form, since the form alone does not capture representational (nor behavioral) adequacy. A cloud parameterization scheme makes sense only with parameter values already assigned and the same holds for a many-parameter density functional. Before the process of adjustment, the mere form of the functional does not offer anything to be called adequate or inadequate. In simulation models, as we have seen, (predictive) success and adaptation are entangled.

The separation of verification and validation thus cannot be fully maintained in practice. It is not possible to first verify that a simulation model is 'right' before tackling the 'external' question whether it is the right model. Performance tests hence become the main handle for confirmation. This is a version of confirmation holism that points toward the limits of analysis. This does not lead to a complete conceptual breakdown of verification and validation. Rather, holism comes in degrees[13] and is a pernicious tendency that undermines the verification-validation divide.[14]

Finally, we come back to the analogy, or rather dis-analogy between computer and clockwork. In an important sense, computers are not amplifiers, i.e. they are not analogous to gigantic clockworks. They do not (simply) amplify the force of mathematical modeling that has got stuck

[13] I thank Rob Muir for pointing this out to me.

[14] My conclusion about the inseparability of verification and validation is in good agreement with Winsberg's more specialized claim in (2010) where he argues about model versions that evolve due to changing parameterizations, which has been criticized by Morrison (2015). As far as I can see, her arguments do not apply to the case made in this paper, which rests on a tendency toward holism, rather than a complete conceptual breakdown.

in too demanding operations. Rather, computer simulation is profoundly *changing* the setting of how mathematics is used.

In the present paper I questioned the rational picture of design. Also Brooks did this when he observed that Pahl and Beitz had to include more and more steps to somehow capture an unwilling and complex practice of design, or when he refers to Donald Schön who criticized a one-sided "technical rationality" that underlies the Rational Model (Brooks 2010, chapter 2). However, my criticism works, if you want, from 'within'. It is the very methodology of simulation modeling, and how it works in practice, that challenges the rational picture by making modularity erode.

The challenge to the rational picture has quite fundamental ramification because this picture influenced so many ways we conceptualize our world. I will spare the philosophical discussion of how simulation modeling is challenging our concept of mathematization and with it our picture of scientific rationality for another paper. Just let me mention the philosophy of mind as one example. How we are inclined to think about mind today is deeply influenced by the computer and by our concept of mathematical modeling. Jerry Fodor has defended a most influential thesis that mind is composed of information-processing devices that operate largely separately (Fodor 1983). Consequently, re-thinking how computer models are related to modularity invites to re-thinking the computational theory of the mind.

I would like to thank …

References

Agre, Philip E., Hierarchy and History in Simon's "Architecture of Complexity", journal of the learning sciences, 3, 2003, 413-426.

Brooks, Frederic P, The Design of Design. Boston, MA: Addison-Wesley, 2010.

Clark, Andy, The Kludge in the Machine, in: Mind and Language 2(4), 1987, 277-300.

Fillion, Nicolas, 2017, The Vindication of Computer Simulations, in Lenhard, J., and Carrier, M. (eds.), *Mathematics as a Tool*, Boston Studies in History and Philosophy of Science, forthcoming.

Fodor, Jerry: *The Modularity of Mind*, 1983, MIT Press, Cambridge, MA.

Foote, Brian und Joseph Yoder, *Pattern Languages of Program Design 4* (= *Software Patterns*. 4). Addison Wesley, 1999.

Frigg, Roman and Julian Reiss, The Philosophy of Simulation. Hot New Issues or Same Old Stew?, in: *Synthese*, 169(3), 593-613, 2009.

Gabriel, Richard P.: *Patterns of Software. Tales From the Software Community*, New York and Oxford: Oxford University Press, 1996.

Gramelsberger, Gabriele und Johann Feichter (eds.): *Climate Change and Policy. The Calculability of Climate Change and the Challenge of Uncertainty*, Heidelberg: Springer 2011.

Hasse, Hans, and Lenhard, J. (2017), On the Role of Adjustable Parameters, in Lenhard, J., and Carrier, M. (eds.), *Mathematics as a Tool*, Boston Studies in History and Philosophy of Science, forthcoming.

Humphreys, Paul, The Philosophical Novelty of Computer Simulation Methods, *Synthese*, 169 (3):615 - 626 (2009).

Jacob, Francois, *The Possible and the Actual*, Seattle: University of Washington Press, 1994.

Lenhard, Johannes, *Disciplines, Models, and Computers: The Path To Computational Quantum Chemistry*, Studies in History and Philosophy of Science Part A, 48 (2014), 89-96.

Lenhard, Johannes and Eric Winsberg, *Holism, Entrenchment, and the Future of Climate Model Pluralism*, in: Studies in History and Philosophy of Modern Physics, 41, 2010, 253-262.

Mauritsen, Thorsten, Bjorn Stevens, Erich Roeckner, Traute Crueger, Monika Esch, Marco Giorgetta, Helmuth Haak, Johann Jungclaus, Daniel Klocke, Daniela Matei, Uwe Mikolajewicz, Dirk Notz, Robert Pincus, Hauke Schmidt, and Lorenzo Tomassini, Tuning the climate of a global model, *Journal of Advances in Modeling Earth Systems*, 4, 2012.

Morrison, Margaret, *Reconstructing Reality. Models, Mathematics, and Simulations.* New York: Oxford University Press, 2015.

Oberkampf, William L., and Roy, Christopher J., *Verification and Validation in Scientific Computing*. Cambridge, MA: Cambridge University Press, 2010.

Oberkampf, William L. and Trucano, T.G., *Validation Methodology in Computational Fluid Dynamics*, American Institute for Aeronautics and Astronautics, 2000 – 2549, 2000.

Pahl, G. and Beitz, W. 1984. *Engineering Design: A Systematic Approach*. Revised editions in 1996, 2007. Berlin: Springer.

Parker, Wendy, Values and Uncertainties in Climate Prediction, revisited, *Studies in History and Philosophy of Science* 2013.

Perdew, J. P., Ruzsinsky, A., Tao, J., Staroverov, V., Scuseria, G., & Csonka, G. (2005). Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *The Journal of Chemical Physics*, 123.

Royce, Wynston, Managing the Development of Large software Systems. *Proceedings of IEEE WESCON* 26 (August), 1970, 1–9.

Simon, Herbert A., *The Sciences of the Artificial*, Cambridge, MA: The MIT Press, 1969.

Smith, Leonard A., What Might We learn From Climate Forecasts?, in: *Proceeedings of the National Academy of Sciences USA*, 4(99), 2002, 2487-2492.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.), *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel*

on Climate Change, 2007. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Stainforth, D.A., Downing, T.E., Washington, R. and New, M. (2007) Issues in the interpretation of climate model ensembles to inform decisions, *Philosophical Transactions of the Royal Society*, Volume 365, Number 1857, 2145-2161.

Wimsatt, William C., *Re-Engineering Philosophy for Limited Beings. Piecewise approximations to reality*, Cambridge, MA and London, England: Harvard University Press, 2007.

Winograd, Terry und F. Flores, *Understanding Computers and Cognition*, Reading, MA: Addison-Wesley, ^{5}1991.

Winsberg, Eric, *Science in the Age of Computer Simulation*, Chicago, Ill.: University of Chicago Press, 2010.

PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association -370-

Accuracy, conditionalization, and probabilism

Peter J. Lewis, University of Miami
Don Fallis, University of Arizona

March 3, 2016

Abstract

Accuracy-based arguments for conditionalization and probabilism appear to have a significant advantage over their Dutch Book rivals. They rely only on the plausible epistemic norm that one should try to decrease the inaccuracy of one's beliefs. Furthermore, it seems that conditionalization and probabilism follow from a wide range of measures of inaccuracy. However, we argue that among the measures in the literature, there are some from which one can prove conditionalization, others from which one can prove probabilism, and none from which one can prove both. Hence at present, the accuracy-based approach cannot underwrite both conditionalization and probabilism.

A central concern of epistemology is uncovering the rational constraints on an agent's credences, both at a time and over time. At a time, it is typically maintained that an agent's credences should conform to the probability axioms, and over time, it is often maintained that an agent's credences should conform to conditionalization. How could such norms be justified? The traditional approach is to show that if your credences violate these norms, then there is a set of bets, each of which you consider fair, but which collectively are such that if you accept them all you will lose money whatever happens. Since you do not want to be a "money pump", you should adopt coherent credences. However, this *Dutch book* strategy rests on controversial assumptions concerning prudential rationality and its connection to epistemic rationality.

The prudential elements may not be essential to the Dutch book approach (Vineberg 2012). But even so, it would be better to be able to derive probabilism and conditionalization from a clearly epistemic basic norm. A more

recent approach seeks to do precisely that: to derive probabilism and conditionalization from the intuitive epistemic norm that you should endeavor to make your credences as accurate—as close to the truth—as possible. Drawing on the work of Joyce (1998; 2009), Greaves and Wallace (2006) and Predd et al. (2009), Pettigrew (2013) argues that the accuracy-based approach vindicates both probabilism and conditionalization. We argue that this conclusion is too strong: at present, the accuracy-based approach can vindicate *either* conditionalization *or* probabilism, but not both.

Our argument turns on the features of various proposed measures of accuracy. The accuracy-based approach is predicated on the assumption that the accuracy of your credences can be measured. Pettigrew (2013, 905) argues that it is a strength of the accuracy-based approach that conditionalization and probabilism follow from a wide range of measures, so that it doesn't matter which measure is used to assess the accuracy of an agent's credences. Our counter-argument is that it does matter: of the known measures, some vindicate conditionalization, and some vindicate probabilism, but there is no known measure of inaccuracy from which both conditionalization and probabilism can be derived.

1 Accuracy and conditionalization

First, let us briefly run through the argument via which conditionalization and probabilism are claimed to follow from considerations of accuracy, starting with conditionalization. Suppose you have credences $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ in propositions $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, where the propositions form a partition, i.e. they are exhaustive and mutually exclusive, so that exactly one of them is true. The accuracy approach takes it that your primary epistemic goal is having credences that are as accurate as possible, where complete accuracy is a credence of 1 in the true proposition and a credence of 0 in each of the false propositions. The closer your credences are to complete accuracy, the better.

For this epistemic goal to make sense, we need a measure of closeness. In what follows we will discuss several such measures, expressed as measures of *inaccuracy*: the larger the measure, the further your credences are from the truth. Hence your goal is to minimize the value of this inaccuracy measure. By far the dominant measure in the literature is the quadratic rule or Brier rule, which takes the square of the difference between your credence in each

proposition and its truth value, and sums the results. So for a partition, if $I_i(\mathbf{b})$ is the inaccuracy of credences $\mathbf{b}$ when proposition X_i is true, then the Brier rule can be expressed as follows:[1]

Simple Brier rule: $I_i(\mathbf{b}) = (1 - b_i)^2 + \sum_{j \neq i} b_j^2$.

The Brier rule has been defended by epistemologists (Joyce 2009, 290; Leitgeb and Pettigrew 2010, 219), and is frequently cited as the prime example of an inaccuracy measure (Greaves and Wallace 2006, 627; Pettigrew 2013, 899).

Suppose you obtain evidence E that is consistent with some but not all of the propositions $\mathbf{X}$. How should you distribute your credence over the remaining propositions? If your goal is to minimize your inaccuracy, presumably the best you can do is to minimize your *expected* inaccuracy given your prior credences $\mathbf{b}$. So suppose that after you learn E, you shift your credence in proposition X_i from b_i to x. If X_i is true, the contribution of this new credence to your overall inaccuracy is $(1 - x)^2$, and if X_i is false, the contribution is x^2. Given your prior credences $\mathbf{b}$, you judge that the chance that X_i is true is b_i, and the chance that X_i is false is $\sum_{E-i} b_j$, where the notation $E - i$ indicates that the sum is over all propositions consistent with E except X_i. That is, the total contribution C of this new credence to your expected inaccuracy is given by:

$$C = (1 - x)^2 b_i + x^2 \sum_{E-i} b_j.$$

Your goal is to minimize C. So consider where $dC/dx = 0$:

$$\begin{aligned} \frac{dC}{dx} &= -2(1 - x) b_i + 2x \sum_{E-i} b_j \\ &= -2b_i + 2x \sum_{E} b_j, \end{aligned}$$

where the sum in the last line is now over all propositions consistent with E. This expression is zero when

$$x = \frac{b_i}{\sum_E b_j}.$$

[1] We call the version of the Brier rule applicable to a partition the *simple* Brier rule only for ease of reference (and similarly for the simple log rule and simple spherical rule to be introduced later).

But note that this value for x is just your prior credence in X_i conditional on E:

$$c(X_i|E) = \frac{c(X_i \wedge E)}{c(E)} = \frac{b_i}{\sum_E b_j}.$$

That is, conditionalizing on E minimizes your expected inaccuracy.[2] So if your epistemic goal is to minimize inaccuracy, you should conditionalize on new evidence.

Greaves and Wallace (2006) generalize this proof to cover measures of inaccuracy other than the Brier rule. In particular, they show that conditionalization minimizes expected inaccuracy for any measure of inaccuracy $I_i(\mathbf{b})$ satisfying *strict propriety*:

Strict propriety: For any distinct probabilistic credences $\mathbf{b}$ and $\mathbf{b}'$, $\sum_i b_i I_i(\mathbf{b}) < \sum_i b_i I_i(\mathbf{b}')$.

Strict propriety says that the expected inaccuracy of your current credences $\mathbf{b}$ is lower than the expected inaccuracy of any alternative credences $\mathbf{b}'$ you might adopt, where the expectation is calculated according to your current credences. If it fails, then the injunction to minimize inaccuracy makes your beliefs pathologically unstable: you can lower your expected inaccuracy by shifting your credences, even in the absence of new evidence. Hence strict propriety serves as a reasonable constraint on measures of inaccuracy. The Brier rule is strictly proper, as are several other proposed inaccuracy measures to be discussed below.

Greaves and Wallace begin by introducing some terminology. They say that a set of credences $\mathbf{b}$ *recommends* a set of credences $\mathbf{b}'$ iff the expected inaccuracy of $\mathbf{b}'$ is at least as low as the expected inaccuracy of $\mathbf{b}$, where the expectation is calculated using credences $\mathbf{b}$:

Recommendation: $\mathbf{b}$ recommends $\mathbf{b}'$ iff $\sum_i b_i I_i(\mathbf{b}) \geq \sum_i b_i I_i(\mathbf{b}')$

Note that if the inaccuracy measure $I_i(\mathbf{b})$ satisfies strict propriety, then $\mathbf{b}$ only recommends itself.

They further define *quasi-conditionalization* as a belief updating rule that stipulates that your credences on learning E should be some set *recommended* by your prior credences conditional on E. They then prove

[2] This proof is a simplified version of the one in Leitgeb and Pettigrew (2010).

that quasi-conditionalization is always optimal: whatever measure of inaccuracy you choose, strictly proper or not, the expected inaccuracy of quasi-conditionalizing is at least as low as the expected inaccuracy of any other updating rule. Then if your measure of inaccuracy is strictly proper, conditionalization itself is optimal, since for strictly proper measures, credences only recommend themselves. In fact, since the inequality in strict propriety is *strict*, conditionalization is strictly better than any other updating rule: it uniquely minimizes expected inaccuracy. As Pettigrew (2013, 905) notes, this is a strong result: any inaccuracy measure satisfying strict propriety can be used to vindicate conditionalization, and strict propriety is a constraint we would expect any reasonable inaccuracy measure to obey anyway.

2 Accuracy and probabilism

Now let us turn to the arguments that your credences at a time should obey the probability axioms. So far, we have been assuming that the propositions we are interested in form a partition. But the probability axioms include constraints on your credences in disjunctions, and to model such constraint we need to allow that more than one of the propositions you are considering can be true. To that end, suppose that you have credences $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ in propositions $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, where now the set of propositions forms a Boolean algebra, i.e. it is closed under negation and disjunction. So now we can no longer model a possible world simply as an index (picking out the unique true proposition); instead, we need to label each proposition separately as either true or false. That is, a possible world is specified by $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots \omega_n)$, where $\omega_i = 1$ when X_i is true and $\omega_i = 0$ when X_i is false. In this context, the Brier rule can be rewritten as follows:

Symmetric Brier rule: $I(\boldsymbol{\omega}, \mathbf{b}) = \sum_i (b_i - \omega_i)^2$.

As before, the inaccuracy of your beliefs according to the Brier rule is given by the sum of the squares of the distance of each belief from the relevant truth value. That is, the Brier rule is *symmetric*, in the sense that distance from the truth for a true proposition plays the same role as distance from falsity plays for a false proposition. This property will be important later.

The general strategy for defending probabilism based on accuracy goes as follows. Suppose that your current credences are incoherent—that is, they

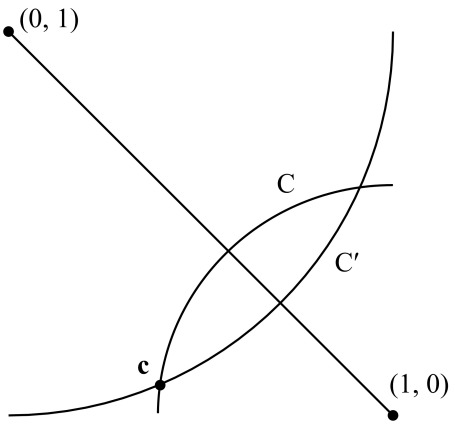


Figure 1: De Finetti's construction for a two-element partition (Joyce 1998, 582).

violate the probability axioms. Then one can appeal to a measure of inaccuracy to show that there are coherent credences that *dominate* your current credences—that are more accurate than your current credences whatever the truth values of the propositions concerned. If your goal is to minimize inaccuracy, this gives you a clear reason to avoid incoherent credences: there are always coherent credences that are more accurate, whatever the world is like.

De Finetti (1974, 87) constructs a dominance argument of this kind based on the Brier rule.[3] For illustration, consider the simple case of a proposition and its negation: that is, the propositions under consideration are just $(X, \neg X)$. In this case the space of possible credences forms a plane, as shown in figure 1: your credence in X is the horizontal coordinate, and your credence in $\neg X$ is the vertical coordinate. The two possible worlds are represented by the points $(1, 0)$ and $(0, 1)$, and your credences obey the probability axioms if and only if they lie on the straight line that connects these two points, since along this line your credences in X and $\neg X$ sum to 1.

Suppose that your credences are incoherent: they are represented by a point $\mathbf{c} = (c_1, c_2)$ that lies *off* this diagonal. And suppose first that the

[3] As Joyce (1998, 580) notes, de Finetti sets up this argument in terms of bets. However, as Pettigrew (2013, 901) points out, it can be redescribed as an accuracy-based argument.

actual world is represented by the bottom-right corner $(1, 0)$—i.e. X is true and $\neg X$ is false. Then the inaccuracy of your credences according to the Brier rule is $I(\boldsymbol{\omega}, \mathbf{c}) = (1 - c_1)^2 + (c_2)^2$. Note that this is just the square of the Euclidean distance between (c_1, c_2) and $(1, 0)$. That is, every point on the circle segment C has the same inaccuracy as $\mathbf{c}$, and every point between C and $(1, 0)$ has a lower inaccuracy. Now suppose instead that the actual world is represented by the top-left corner $(0, 1)$—i.e. X is false and $\neg X$ is true. Then the inaccuracy of your credences is $I(\boldsymbol{\omega}, \mathbf{c}) = (c_1)^2 + (1 - c_2)^2$—the square of the Euclidean distance between (c_1, c_2) and $(0, 1)$. That is, every point on the circle segment C' has the same inaccuracy as $\mathbf{c}$, and every point between C and $(0, 1)$ has a lower inaccuracy.

Consider the area enclosed by the circle segments C and C'. The credences represented by the points in this area have a lower inaccuracy than $\mathbf{c}$ if X is true and $\neg X$ false, and a lower inaccuracy than $\mathbf{c}$ if X is false and $\neg X$ true. That is, they have a lower inaccuracy whatever the world is like. And this area includes part of the diagonal that represents coherent credences. So for any incoherent set of credences, there is a coherent set that is less inaccurate whatever the world is like. In this simple case, accuracy gives you a motive to adopt coherent credences.

In the general case, the space of possible credences is n-dimensional, where there are n propositions in the Boolean algebra. Each possible assignment of truth values to the n propositions is represented by a point in this space, and the set of coherent credences consists of these points plus the points on the straight lines that connect them, the points on the straight lines that connect those latter points, and so on. This set is called the *convex hull* V^+ of the possible truth value assignments V. Via a generalization of the construction of figure 1, de Finetti shows that if your credences are represented by a point that lies outside V^+, then there are points in V^+ that are more accurate (according to the Brier rule) whichever point in the space represents the actual truth values of the propositions. Hence if you have incoherent credences, there are always coherent credences with a lower inaccuracy as measured by the Brier rule.

Predd et al. (2009) generalize this proof strategy to cover a wider range of inaccuracy measures. Their proof relies on two assumptions. The first is additivity:

Additivity: $I(\boldsymbol{\omega}, \mathbf{b})$ can be expressed as $\sum_i s(\omega_i, b_i)$, where s is a continuous function of your credence in proposition X_i and its truth value.

Additivity states that the inaccuracy of your beliefs in a set of propositions is just the sum of your inaccuracies in the propositions taken individually—that is, $s(\omega_i, b_i)$ is the inaccuracy of your belief in proposition X_i, and $I(\boldsymbol{\omega}, \mathbf{b})$ is just the sum of these inaccuracies for all the propositions you are considering. Note that it also contains the requirement that the inaccuracy measure should be continuous. The Brier rule is obviously additive, since it is expressed as a sum over propositions.

The second assumption is a version of strict propriety. For an additive inaccuracy measure, strict propriety can be expressed in terms of your inaccuracy function for a single proposition $s(b_i, \omega_i)$ as follows:

Strict propriety (for an additive measure): $b_i s(x, 1) + (1 - b_i)s(x, 0)$ is uniquely minimized at $x = b_i$.

Predd et al. (2009) prove that any additive, strictly proper inaccuracy measure entails probabilism. De Finetti's construction appeals to the natural distance measure implicit in the Brier rule—the Euclidean distance between two points in the space of your possible credences. But in the current case we have no explicit measure of inaccuracy, so Predd et al. appeal to a generalized "distance" measure[4] called the Bregman divergence, defined for a strictly convex function $\Phi(\mathbf{x})$ as $d_\Phi(\mathbf{y}, \mathbf{x}) = \Phi(\mathbf{y}) - \Phi(\mathbf{x}) - \nabla\Phi(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x})$. They show that if the inaccuracy measure $s(b_i, \omega_i)$ for a single proposition X_i is strictly proper, then the function $\varphi(b_i) = -b_i s(b_i, 1) - (1 - b_i)s(b_i, 0)$ is strictly convex. In terms of this function, Predd et al. show that for any additive, strictly proper inaccuracy measure, $I(\boldsymbol{\omega}, \mathbf{b}) = d_\Phi(\boldsymbol{\omega}, \mathbf{b})$, where $\Phi(\boldsymbol{\omega}) = \sum_i \varphi(\omega_i)$ and $\Phi(\mathbf{b}) = \sum_i \varphi(b_i)$.

The set of coherent credences forms a closed, convex subspace V^+ of the space of all possible credences. It is a fact from the theory of Bregman divergences that for any point $\mathbf{c}$ outside V^+, there is a unique point $\mathbf{c}^*$ in V^+ such that $d_\Phi(\mathbf{c}^*, \mathbf{c}) \leq d_\Phi(\mathbf{y}, \mathbf{c})$ for all $\mathbf{y}$ in V^+. That is, $\mathbf{c}^*$ is the unique closest point in V^+ to $\mathbf{c}$, using the Bregman divergence as a distance measure. It is a further fact that $d_\Phi(\mathbf{y}, \mathbf{c}^*) \leq d_\Phi(\mathbf{y}, \mathbf{c}) - d_\Phi(\mathbf{c}^*, \mathbf{c})$ for all $\mathbf{y}$ in V^+ and $\mathbf{c}$ outside V^+. Note in particular that V^+ contains every possible world $\boldsymbol{\omega}$, since a consistent truth value assignment is also a coherent set of credences. So setting $\mathbf{y} = \boldsymbol{\omega}$, we have $d_\Phi(\boldsymbol{\omega}, \mathbf{c}^*) \leq d_\Phi(\boldsymbol{\omega}, \mathbf{c}) - d_\Phi(\mathbf{c}^*, \mathbf{c})$. Since d_Φ is a positive-valued function, $d_\Phi(\mathbf{c}^*, \mathbf{c}) > 0$, so $d_\Phi(\boldsymbol{\omega}, \mathbf{c}^*) < d_\Phi(\boldsymbol{\omega}, \mathbf{c})$, and hence

[4] The reason for the scare quotes is that the Bregman divergence is not symmetric, and distance measures are typically symmetric.

$I(\boldsymbol{\omega}, \mathbf{c}^*) < I(\boldsymbol{\omega}, \mathbf{c})$. That is, for any incoherent set of credences $\mathbf{c}$, there is a coherent set $\mathbf{c}^*$ that is less inaccurate than $\mathbf{c}$ in every possible world.

As Pettigrew (2013, 905) notes, this is a strong result: any inaccuracy measure satisfying strict propriety and additivity can be used to vindicate probabilism, and while additivity is perhaps not forced on us in the way that strict propriety is, it is certainly intuitive. As we shall see, there are several available measures satisfying additivity and strict propriety, so it initially looks like the accuracy-based program can justify both probabilism and conditionalization based on minimal premises. Our purpose in this paper is to argue that matters are not so straightforward.

3 Measures of inaccuracy

Let us return to the argument for conditionalization. This argument restricts inaccuracy measures to those that are strictly proper. Note that strict propriety is only a condition on *expected* inaccuracy. But expected inaccuracy is calculated on the basis of the *actual* inaccuracy that the measure in question ascribes to credences, and presumably there are a number of constraints any such measure must obey if it is to genuinely measure epistemic inaccuracy rather than something else. For example, if one of your credences shifts towards the truth, while your other credences stay the same, then clearly your actual inaccuracy should decrease. We wish to focus on one such constraint.

The constraint can be motivated by thinking about *elimination cases*. Suppose you are considering a set of mutually exclusive and exhaustive propositions, and suppose that your credences are coherent and that you conditionalize on evidence. You acquire some evidence that eliminates one false proposition—your credence in it becomes zero—but is uninformative regarding the other hypotheses—your credences in them remain in the same proportions. How does this affect the accuracy of your credences?

It seems obvious that your beliefs have become more accurate. If you believe that Tom, Dick or Harry might be the murderer (when in fact Tom did it), and you eliminate Harry while learning nothing about Tom or Dick, then you have made epistemic progress towards the truth, or at least away from falsity. It is true that your credence in the false proposition "Dick did it" goes up, but only by the same proportion that your credence in the true proposition "Tom did it" goes up.

Unfortunately, the simple Brier rule does not always concur. Let X_1 be

"Tom did it", X_2 be "Dick did it", and X_3 be "Harry did it", where unknown to you X_1 is true. Suppose that your initial credences in (X_1, X_2, X_3) are $\mathbf{b} = (1/7, 3/7, 3/7)$. Then according to the simple Brier rule, your initial inaccuracy is $54/49 = 1.10$. Now suppose you acquire some evidence that eliminates X_3, but is uninformative regarding X_1 and X_2. That is, your credence in X_3 becomes 0 and your credences in X_1 and X_2 stay in the same proportions, so that your final credences are $\mathbf{b}^* = (1/4, 3/4, 0)$. Then according to the simple Brier rule, your final inaccuracy is $18/16 = 1.13$. That is, the Brier rule erroneously says that the inaccuracy of your beliefs has gone up.

For a measure to genuinely measure the actual inaccuracy of your beliefs, it should not be susceptible to counterexamples of this kind; it should count elimination cases as epistemically positive. That is, measures of inaccuracy should obey the following principle:

M: For coherent credences over a partition, if $\mathbf{b}$ assigns a zero credence to some false proposition to which $\mathbf{b}'$ assigns a non-zero credence, and credences in the remaining propositions stay in in the same ratios, then $\mathbf{b}$ is epistemically better than $\mathbf{b}'$.

The simple Brier rule, as the example shows, violates M, and hence does not plausibly measure the actual inaccuracy of your beliefs.[5]

Fortunately, though, there are alternative inaccuracy measures for partitions we can appeal to. The two most frequently mentioned are the simple log rule and the simple spherical rule:

Simple log rule: $I_i(\mathbf{b}) = -\ln b_i$

Simple spherical rule: $I_i(\mathbf{b}) = 1 - b_i/\sqrt{\sum_j b_j^2}$.

As before, $I_i(\mathbf{b})$ is the inaccuracy of credences $\mathbf{b}$ when proposition X_i is true. Both of these measures satisfy M, and hence are not susceptible to elimination counterexamples.[6] Hence each can plausibly be claimed to measure epistemic inaccuracy. Furthermore, each is strictly proper, and so each can be used to

[5] One might reasonably think that acceptable measures of accuracy should obey a stronger principle than M; see (*reference removed*).

[6] This is trivial for the log rule, and easily proven for the spherical rule. See (*reference removed*).

underwrite conditionalization via the above argument strategy. So there are some inaccuracy measures that vindicate conditionalization, but not all strictly proper measures do so. In particular, the simple Brier rule cannot be used to vindicate conditionalization.

But what about probabilism? The simple log rule and simple spherical rule are not applicable to a Boolean algebra, and so cannot be used to prove probabilism as they stand. Perhaps the most straightforward way to generalize them is simply to sum the contribution given by the simple rule for each true proposition in the Boolean algebra, while ignoring the false propositions in the algebra:

Asymmetric log rule: $I(\mathbf{b}, \boldsymbol{\omega}) = \sum_i F(\omega_i, b_i)$, where $F(0, b_i) = 0$ and $F(1, b_i) = -\ln b_i$.

Asymmetric spherical rule: $I(\mathbf{b}, \boldsymbol{\omega}) = \sum_i F(\omega_i, b_i)$, where $F(0, b_i) = 0$ and $F(1, b_i) = 1 - b_i / \sqrt{\sum_j b_j^2}$.

Both these rules are asymmetric, in the sense that inaccuracy is calculated differently for true and false propositions. These rules satisfy principle M: for coherent credences, if your credence in a false proposition goes down and your remaining credences stay in the same ratios, then your credence in each true proposition goes up, and so your inaccuracy according to the relevant asymmetric rules goes down. Hence the asymmetric log and spherical rules are immune from elimination counterexamples.

But these rules do not satisfy the combination of additivity and strict propriety required for the proof of probabilism. The asymmetric spherical rule is not additive: $F(1, b_i)$ is not a function of b_i alone. The asymmetric log rule is additive, but it is not strictly proper in the required sense: $F(1, b_i)$ is strictly proper, but $F(0, b_i)$ is not. Indeed, it is straightforward to show directly that these rules cannot be used as the basis of a dominance argument for probabilism. Consider, for example, a two element partition, and the incoherent credence assignment $(1, 1)$. The asymmetric log rule counts these incoherent credences as *perfectly* accurate (since the credence in the false proposition is ignored), so no coherent credences can dominate them. According to the asymmetric spherical rule, multiplying all credences by a constant has no effect on inaccuracy, so this assignment has the same inaccuracy as the coherent credence assignment $(1/2, 1/2)$. If coherent assignments cannot be dominated, then neither can the initial incoherent assignment.

But if coherent assignments *can* be dominated then the dominance proof of probabilism fails anyway.

So the asymmetric versions of the log rule and the spherical rule cannot be used to prove probabilism. But for a Boolean algebra, the log rule and the spherical rule are usually given a formulation that is symmetric between truth and falsity:

Symmetric log rule: $I(\boldsymbol{\omega}, \mathbf{b}) = \sum_i - \ln |(1 - \omega_i) - b_i|$

Symmetric spherical rule: $I(\boldsymbol{\omega}, \mathbf{b}) = \sum_i 1 - \dfrac{|(1 - \omega_i) - b_i|}{\sqrt{b_i^2 + (1 - b_i)^2}}$

(see e.g. Joyce 2009, 275). These measures are additive, and each term in the sum is individually strictly proper, so they can each be used to prove probabilism via the proof of Predd et al.

But unfortunately, in their symmetric forms all three rules—Brier, log and spherical—are subject to elimination counterexamples. For the Brier rule, the counterexample is the same as before, since the symmetric Brier rule reduces to the simple Brier rule when applied to a partition.[7] That is, consider a credence shift from $\mathbf{b} = (1/7, 3/7, 3/7)$ to $\mathbf{b}^* = (1/4, 3/4, 0)$ when X_1 is true. According to the symmetric Brier rule, your initial inaccuracy is 1.10, and your final inaccuracy is 1.13, so your inaccuracy goes up. And this example works equally well against the symmetric spherical rule: according to this rule, your initial inaccuracy is 1.24 and your final inaccuracy is 1.37, so your inaccuracy goes up. This particular counterexample does not work against the symmetric log rule, but a similar one does. Suppose your initial credences are $\mathbf{b} = (1/13, 6/13, 6/13)$, and your final credences are $\mathbf{b}^* = (1/7, 6/7, 0)$. Then according to the symmetric log rule your initial inaccuracy is 3.80, and your final inaccuracy is 3.89: your inaccuracy goes up. Hence the symmetric measures all violate principle M, and so none of them can be used to prove conditionalization.

[7] Strictly, applying these rules to a Boolean algebra requires including credences in the negations $\neg X_1$, $\neg X_2$ and $\neg X_3$, plus the tautology $X_1 \vee X_2 \vee X_3$ and the contradiction $\neg(X_1 \vee X_2 \vee X_3)$. But for coherent credences the inaccuracies of the tautology and the contradiction are zero, and for symmetric rules the inaccuracy of $\neg X_i$ is the same as that of X_i, so the inaccuracy calculated over the entire Boolean algebra is simply twice the inaccuracy over the partition (X_1, X_2, X_3).

4 The extent of the problem

Let us sum up. The simple Brier rule cannot be used to prove conditionalization, but the simple log and spherical rules can. The obvious generalizations of the simple log and spherical rules to a Boolean algebra—the asymmetric log and spherical rules—cannot be used to prove probabilism. The symmetric Brier, log and spherical rules can be used to prove probabilism, but none of them underwrites conditionalization. So we have found no measure that can be used to prove both conditionalization *and* probabilism.

Could there be such a measure? Perhaps, although it is worth noting that one can prove that *any* inaccuracy measure that satisfies additivity, strict propriety and a plausible symmetry principle is subject to elimination counterexamples. The symmetry principle is precisely the one discussed above—that the inaccuracy measure treats truth the same as falsity, in the sense that it is a function of the distance between each credence and its respective truth value. For an additive inaccuracy measure, the symmetry principle can be expressed in terms of the inaccuracy function for a single proposition $s(\omega_i, b_i)$ as follows:

Symmetry: $s(\omega_i, b_i) = s(|1 - \omega_i|, |1 - b_i|)$.

It is certainly highly plausible that this is part of what it means for s to measure your distance from the truth, and as discussed above, the typical Boolean algebra forms of the Brier rule, log rule and spherical rule all satisfy it.

Let us see how this symmetry principle, together with additivity and strict propriety, lead to elimination counterexamples. Consider a single proposition X_i in which your credence is $b_i = 1/2$. According to strict propriety, the quantity $(1/2)s(1, x) + (1/2)s(0, x)$ must be uniquely minimized at $x = 1/2$. In particular, the value of this expression for $x = 1/2$ must be lower than its value for $x = 1$:

$$(1/2)s(1, 1/2) + (1/2)s(0, 1/2) < (1/2)s(1, 1) + (1/2)s(0, 1),$$

and for $x = 0$:

$$(1/2)s(1, 1/2) + (1/2)s(0, 1/2) < (1/2)s(1, 0) + (1/2)s(0, 0).$$

Adding these:

$$s(1, 1/2) + s(0, 1/2) < (1/2)s(1, 1) + (1/2)s(0, 1) + (1/2)s(1, 0) + (1/2)s(0, 0).$$

But by symmetry, $s(1, 1/2) = s(0, 1/2)$, $s(1, 1) = s(0, 0)$ and $s(0, 1) = s(1, 0)$. Substituting:

$$2s(0, 1/2) < s(0, 1) + s(0, 0).$$

Now consider your credences in three exhaustive and mutually exclusive propositions $\mathbf{X} = (X_1, X_2, X_3)$. Consider in particular the credence shift from $\mathbf{m} = (0, 1/2, 1/2)$ to $\mathbf{b} = (0, 1, 0)$ for truth values $\boldsymbol{\omega} = (1, 0, 0)$. By separability, $I(\boldsymbol{\omega}, \mathbf{m}) = s(1, 0) + 2s(0, 1/2)$, and $I(\boldsymbol{\omega}, \mathbf{b}) = s(1, 0) + s(0, 1) + s(0, 0)$. So since $2s(0, 1/2) < s(0, 1) + s(0, 0)$ it follows that $I(\boldsymbol{\omega}, \mathbf{m}) < I(\boldsymbol{\omega}, \mathbf{b})$: your inaccuracy goes up. But the shift from $\mathbf{m} = (0, 1/2, 1/2)$ to $\mathbf{b} = (0, 1, 0)$ is an elimination case: a false proposition is eliminated, and your credences in the remaining hypotheses stay in the same proportions. And lest one worry about the fact that your initial credence in the true proposition is zero, we can modify the example. Consider the credence assignments $\mathbf{m}' = (\delta/(2 + \delta), 1/(2 + \delta), 1/(2 + \delta))$ and $\mathbf{b}' = (\delta/(1 + \delta), 1/(1 + \delta), 0)$. For small δ these are close to $\mathbf{m}$ and $\mathbf{b}$, and hence by the continuity clause of additivity, the inaccuracy of $\mathbf{m}'$ remains lower than that of $\mathbf{b}'$. Again, the transition from $\mathbf{m}'$ to $\mathbf{b}'$ is an elimination case, and now your credence in the true proposition is non-zero.

So elimination counterexamples afflict any inaccuracy measure that satisfies additivity, strict propriety and symmetry. That is, any symmetric measure that satisfies the assumptions of Predd et al.'s proof of probabilism violates principle M, and hence cannot be used to prove conditionalization. Symmetry is not a premise in the Predd argument, so it is possible that an asymmetric measure might allow the derivation of both probabilism and conditionalization. But the only plausible asymmetric measure in the literature is the log rule (Bernardo 1979), and we have seen that the asymmetric log rule does not vindicate probabilism.

5 Conclusion

Pettigrew notes that conditionalization and probabilism follow from a wide range of measures of inaccuracy, and the implication is that it doesn't much matter which measure you pick. But we think it does matter. There are measures that vindicate conditionalization, and there are measures that vindicate probabilism, but nobody has yet identified a measure that vindicates both. Hence the accuracy-based approach does not, as yet, give us the justification we might want for the constraints on our credences.

References

Bernardo, José M. (1979), "Expected information as expected utility", *Annals of Statistics* 7: 686-690.

de Finetti, Bruno (1974), *Theory of Probability*, vol. 1. New York: John Wiley and Sons.

Greaves, Hilary and David Wallace (2006), "Justifying conditionalization: conditionalization maximizes expected epistemic utility", *Mind* 115: 607–32.

Joyce, James M.. (1998), "A nonpragmatic vindication of probabilism", *Philosophy of Science*, 65: 575–603.

Joyce, James M. (2009), "Accuracy and coherence: prospects for an alethic epistemology of partial belief", in F. Huber and C. Schmidt-Petri (eds.), *Degrees of Belief*. Dordrecht: Springer: 263–97.

Leitgeb, Hannes, and Richard Pettigrew (2010), "An objective justification of Bayesianism I: measuring inaccuracy", *Philosophy of Science* 77: 201–35.

Pettigrew, Richard (2013), "Epistemic utility and norms for credence", *Philosophy Compass* 8: 897–908.

Predd, Joel B., Robert Seiringer, Elliott H. Lieb, Daniel N. Osherson, H. Vincent Poor, and Sanjeev R. Kulkarni (2009), "Probabilistic coherence and proper scoring rules", *IEEE Transactions on Information Theory* 55: 4786–4792.

Vineberg, Susan (2012), "Dutch book arguments", in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/spr2012/entries/dutch-book/

Can Typicality Arguments Dissolve Cosmology's Flatness Problem?

C.D. McCoy*

20 February 2016

Abstract

The flatness problem in cosmology draws attention to a surprising fine-tuning of the spatial geometry of our universe towards flatness. Several physicists, among them Hawking, Page, Coule, and Carroll, have argued against the probabilistic intuitions underlying such fine-tuning arguments in cosmology and instead propose that the canonical measure on the phase space of Friedman-Robertson-Walker spacetimes should be used to evaluate fine-tuning. They claim that flat spacetimes in this set are actually typical on this natural measure and that therefore the flatness problem is illusory. I argue that they misinterpret typicality in this phase space and, moreover, that no conclusion can be drawn at all about the flatness problem by using the canonical measure alone.

For several decades now cosmologists have maintained that the old standard model of cosmology, the highly successful hot big bang (HBB) model, suffers from various fine-tuning problems (Dicke and Peebles, 1979; Linde, 1984). They claim that the spacetimes on which the HBB model is based, the Friedman-Robertson-Walker (FRW) spacetimes, require seemingly "special" initial conditions, such that when they are evolved forward in time by the dynamical law of the general theory of relativity (GTR) they yield presently observed cosmological conditions. For example, the flatness problem depends on the existence of special initial conditions in the HBB model which are required to explain the observationally-inferred spatial flatness of the universe. Due to their extreme precision or intuitive "unlikeliness," these initial conditions are thought to be unduly special, such that many cosmologists have felt that the initial conditions themselves are in need of explanation and, moreover, present a significant conceptual problem for the HBB model.

Although physical fine-tuning could be interpreted in a variety of ways, cosmologists typically understand it to mean that observationally-required initial conditions are in some sense unlikely (Smeenk, 2013; McCoy, 2015). In order to substantiate this interpretation, one must show that initial conditions in the HBB model which reproduce present conditions are in fact unlikely. This task presupposes that there is a justifiable way of assessing the likelihoods of cosmological models (Gibbons et al., 1987; Hawking and Page, 1988). Many arguments found in the cosmological literature, however, rely on ad hoc, unjustified likelihood measures. Gibbons et al. (1987) propose a "natural" measure (hence the GHS measure) on the set of FRW spacetimes (with matter contents represented by a scalar field) as a natural and justified way of evaluating likelihoods. The GHS measure is simply the canonical Liouville measure associated with the phase space of FRW spacetimes when GTR is put into a Hamiltonian formulation and in a precise sense "comes for free" with the phase space.

While I would maintain that the GHS measure cannot be successfully used to make arguments about fine-tuning in cosmology quite generally, I argue here only for its inapplicability to the flatness problem. Some

*Eidyn Research Centre, University of Edinburgh, Edinburgh, UK. email: casey.mccoy@ed.ac.uk

authors (Gibbons and Turok, 2008; Carroll and Tam, 2010) have attempted to make probabilistic arguments, in analogy to familiar probabilistic arguments in statistical mechanics, by making the GHS measure into a probability measure. However, as the total measure of the FRW phase space is infinite, there is no canonical choice of probability measure with which to make probabilistic arguments, a point that has been recognized already by some (Hawking and Page, 1988; Schiffrin and Wald, 2012). Accordingly, any justification of a particular probability measure is completely independent of the justification of the GHS measure—in short, these probability measures are not in any substantive sense the GHS measure. On the other hand, one might try to use the GHS measure by itself to make typicality arguments in analogy to typicality arguments in statistical mechanics (Goldstein, 2012). Carroll in particular advocates this approach and, interestingly, claims that the GHS measure alone tells us that almost all spacetimes are spatially flat (Carroll and Tam, 2010; Remmen and Carroll, 2013; Carroll, forthcoming)—that there is in fact no flatness problem (Hawking and Page (1988, 803-4) and Coule (1995, 468) suggest the same). Carroll's claim, however, rests on a subtle mistake in interpreting typicality. I claim, on the contrary, that the GHS measure cannot tell us anything about likelihood without substantive additional assumptions such as those made in statistical mechanics, e.g. a partition of phase space into "macroproperties" or similar. These necessary assumptions, however, are doubtfully justifiable in the cosmological context. Thus I ultimately conclude that the GHS measure cannot be used to clarify the nature of fine-tuning in cosmology.

1 The Gibbons-Hawking-Stewart Measure

An adequate view of what the GHS measure is and can do relies on understanding the details of how it is introduced. For this reason I develop here the measure with considerably more care than other accounts in the literature, which tend to jump straight to a Lagrangian or Hamiltonian formulation of GTR without elucidating the geometrical origin of their variable choices and the relations between physical parameters.

My starting point is the initial value formulation of GTR, in which the "position" initial data of spacetime are represented by the spatial metric h_{ab} on a spacelike Cauchy surface Σ and the "momentum" initial data by the extrinsic curvature π_{ab} (Wald, 1984; Malament, 2012). FRW spacetimes are spacetimes with homogeneous and isotropic spacelike hypersurfaces, so one can foliate the spacetimes by a one-parameter family of these spacelike hypersurfaces Σ_t that are orthogonal to a smooth, future-directed, twist-free, unit timelike field ξ^a on M, where I define $\xi^a = \nabla^a t$.For FRW spacetimes the extrinsic curvature of an initial data surface Σ_t is Hh_{ab}, where H is the so-called Hubble parameter.Thus the initial data for an FRW spacetime are completely represented by two objects: (1) the spatial metric h_{ab} and (2) the Hubble parameter H associated with a spatial hypersurface Σ.

The space of initial data is therefore the product of the set of homogeneous and isotropic Riemannian manifolds Σ (with metric h_{ab}) and the set of (real-valued) Hubble parameters H. Homogeneous and isotropic Riemannian manifolds have constant curvature κ. Complete, connected Riemannian manifolds of constant sectional curvature are called space forms. It is a theorem that every simply-connected three-dimensional space form is isometric to the sphere $S^3(\sqrt{(1/\kappa)})$ if $\kappa > 0$, $\mathbf{R}^3$ if $\kappa = 0$, or the hyperbolic space $H^3(\sqrt{(1/\kappa)})$ if $\kappa < 0$ (Wolf, 2010). The standard metrics on each of these manifolds is understood to be the metric induced on them by embedding them in $\mathbf{R}^4$. Every Σ is therefore isometric to one of these three classes of space forms. Spaceforms of each of the three kinds are moreover homothetic, i.e. they are isometric up to the square of a scale factor a (McCabe, 2004). Accordingly one has the means to represent curvature κ as a function of the scale factor; in particular, for any Σ, $a^2\kappa$ is some constant k. Hence one can set any spatial metric $h_{ab} = a^2\gamma_{ab}$, where γ_{ab} is the standard metric on the appropriate space form. This is useful in the initial value formulation of FRW spacetimes because all time dependence of h_{ab} is thereby located solely in

the scale factor rather than in the radius of curvature of the space form.

The Einstein equation reduces to two constraint equations and two evolution equations in the initial value formulation (Geroch, 1972):

$$\mathcal{R} - (\pi_a{}^a)^2 + \pi_{ab}\pi^{ab} = -16\pi T_{ab}\xi^a\xi^b; \tag{1}$$

$$D_c\pi_a{}^c - D_a\pi_c{}^c = 8\pi T_{mr}h^m{}_a\xi^r; \tag{2}$$

$$\pounds_\xi(\pi_{ab}) = 2\pi_a{}^c\pi_{cb} - \pi^c{}_c\pi_{ab} + \mathcal{R}_{ab} - 8\pi h_a{}^m h^n{}_b(T_{mn} - \frac{1}{2}Th_{mn}); \tag{3}$$

$$\pounds_\xi(h_{ab}) = 2\pi_{ab}, \tag{4}$$

where $\mathcal{R}$ is the Ricci scalar of Σ, $\mathcal{R}_{ab}$ is the Ricci tensor of Σ, and D_a is the derivative operator on Σ. For FRW spacetimes, these equations simplify to the following three (the second equation from above is trivial since π_{ab} does not vary across Σ:

$$\mathcal{R} - 6H^2 = -16\pi\rho; \tag{5}$$

$$\dot{H}h_{ab} = \Big(- H^2 - \frac{4\pi}{3}(\rho + 3p)\Big)h_{ab}; \tag{6}$$

$$\dot{h}_{ab} = 2Hh_{ab}, \tag{7}$$

where ρ is the energy density and p the pressure of the matter. The first two equations are known as the Friedman equations. Since $h_{ab} = a^2\gamma_{ab}$, $\dot{h}_{ab} = 2a\dot{a}\gamma_{ab}$, and $2Hh_{ab} = 2Ha^2\gamma_{ab}$, it follows from the third equation above that

$$H = \frac{\dot{a}}{a}, \tag{8}$$

which is the usual definition of the Hubble parameter H. To simplify matters somewhat and to make contact with the literature, I shall henceforth take the matter contents of spacetime to be a scalar field ϕ in a potential V which evolves according to the coupled Einstein-Klein Gordon equation.[1] Then one has the following equations of motion (Hawking and Page, 1988, 790):

$$\mathcal{R} - 6H^2 = -16\pi\Big(\frac{1}{2}\dot{\phi}^2 + V(\phi)\Big) \tag{9}$$

$$\dot{H} = -H^2 - \frac{8\pi}{3}\Big(\frac{1}{2}\dot{\phi}^2 - V(\phi)\Big) \tag{10}$$

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0, \tag{11}$$

where V' is the derivative of the potential with respect to ϕ.[2] (The third equation can be derived from the previous two, and so is in fact redundant.)

For FRW spacetimes the spatial Ricci scalar is $\mathcal{R} = -6\kappa$. As noted before, one can cast κ in terms of the scale factor and a constant k: $\kappa = k/a^2$. By using the scale factor a to replace κ, one has introduced a constant k which has no physical significance beyond identifying whether the space form is flat, positively-curved, or negatively-curved. One therefore usually takes equivalence classes of curves according to these three cases and chooses $k = +1$, 0, and -1 as representatives. Then one may write $\mathcal{R} = -6k/a^2$, so that one finally has Friedman's equation in its usual form (for a scalar field in a potential):

$$\Big(\frac{\dot{a}}{a}\Big)^2 = \frac{8\pi}{3}\Big(\frac{1}{2}\dot{\phi}^2 + V(\phi)\Big) - \frac{k}{a^2}. \tag{12}$$

[1] The scalar field is meant to be the inflaton, the field that drives inflation in the early universe.

[2] If our interest were solely in assessing the HBB model's fine-tuning, one could do the following analysis for perfect fluid matter contents. The results would be qualitatively similar however, as shown by Carroll and Tam (2010, §4.2).

The foregoing indicates that our FRW initial data h_{ab} and π_{ab} are equivalently representable in the space $\{a, \dot{a}, \phi, \dot{\phi}, k\}$. This space is not the space of initial data, however, since the previous equation is a constraint that must be satisfied by initial data. One must also keep in mind that k is an index for three separate copies of the space $\{a, \dot{a}, \phi, \dot{\phi}\}$. There is no continuous path between the three spaces.

Have identified the relevant spaces for representing FRW space forms, I next put the theory into a Hamiltonian formulation (Wald, 1984, Appendix E) in order to obtain a symplectic structure and, hence, the canonical measure. I begin with the Lagrangian for our theory of FRW spacetimes with a scalar field as the matter contents, where I have re-introduced the lapse function N as a Lagrange multiplier:

$$\mathcal{L} = \sqrt{-g}\Big(\frac{R}{16\pi} + \frac{1}{2N^2}\dot{\phi}^2 - V(\phi)\Big). \tag{13}$$

In terms of the variables I have chosen, this is

$$\mathcal{L} = -\frac{1}{8\pi}(\frac{3}{N}a\dot{a}^2 - 3Na^3\frac{k}{a^2}) + \frac{1}{2N}a^3\dot{\phi}^2 - Na^3V(\phi), \tag{14}$$

in agreement with (Hawking and Page, 1988; Gibbons and Turok, 2008; Carroll and Tam, 2010). The momenta of a and ϕ are

$$p_a \equiv \frac{\partial\mathcal{L}}{\partial\dot{a}} = \frac{-3a\dot{a}}{4\pi N}; \qquad\qquad p_\phi \equiv \frac{\partial\mathcal{L}}{\partial\dot{\phi}} = \frac{a^3\dot{\phi}}{N}. \tag{15}$$

The Hamiltonian on this phase space is

$$\mathcal{H} = p_a\dot{a} + p_\phi\dot{\phi} - \mathcal{L} = N\Big(-\frac{2\pi p_a^2}{3a} + \frac{p_\phi^2}{2a^3} + a^3V(\phi) - a^3\frac{3}{8\pi}\frac{k}{a^2}\Big), \tag{16}$$

from which one recovers (after setting $N = 1$) our constraint (the Friedman equation) as the Hamiltonian constraint C:

$$C \equiv -\frac{2\pi p_a^2}{3a} + \frac{p_\phi^2}{2a^3} + a^3V(\phi) - a^3\frac{3}{8\pi}\frac{k}{a^2} = 0. \tag{17}$$

The phase space γ of our system is thus the four-dimensional space $\{a, p_a, \phi, p_\phi\}$ equipped with the canonical symplectic form

$$\omega_{p_a,a,p_\phi,\phi} = \mathrm{d}p_a \wedge \mathrm{d}a + \mathrm{d}p_\phi \wedge \mathrm{d}\phi. \tag{18}$$

The dynamically accessible phase space points are constrained to be on the three-dimensional hypersurface C. Thus it would be inappropriate to use ω for constructing a canonical volume measure on phase space. One can, however, pull the symplectic form back onto the constraint surface by first solving the constraint for p_ϕ:[3]

$$p_\phi = a^3\Big(\frac{4\pi}{3}\frac{p_a^2}{a^4} + \frac{3}{4\pi}\frac{k}{a^2} - 2V(\phi)\Big)^{1/2}. \tag{19}$$

Following Carroll and Tam, I also switch coordinates from p_a to H, so that

$$p_\phi = a^3\Big(\frac{3}{4\pi}(H^2 + k/a^2) - 2V(\phi)\Big)^{1/2} \tag{20}$$

[3]The scalar field can have positive or negative momentum, so strictly speaking there should be a ± in the following equation. The reader is welcome to annotate the equations that follow.

and

$$\mathrm{d}p_a = -\frac{3}{4\pi}\Big(2aH\mathrm{d}a + a^2\mathrm{d}H\Big). \tag{21}$$

The differential of p_ϕ is then

$$\mathrm{d}p_\phi = \frac{(3/4\pi)a^3H\mathrm{d}H - a^3V'\mathrm{d}\phi + 6a^2((3H^2 + 2k/a^2)/8\pi - V)\mathrm{d}a}{((3/4\pi)(H^2 + k/a^2) - 2V)^{1/2}}. \tag{22}$$

Substituting these into ω then gives the pullback of the symplectic form onto C. The result is the following (pre-symplectic) differential form:

$$\omega_{a,H,\phi} = \Theta_{Ha}(\mathrm{d}H \wedge \mathrm{d}a) + \Theta_{H\phi}(\mathrm{d}H \wedge \mathrm{d}\phi) + \Theta_{a\phi}(\mathrm{d}a \wedge \mathrm{d}\phi), \tag{23}$$

where

$$\Theta_{Ha} = -\frac{3}{4\pi}a^2; \tag{24}$$

$$\Theta_{H\phi} = \frac{(3/4\pi)a^3H}{((3/4\pi)(H^2 + k/a^2) - 2V)^{1/2}}; \tag{25}$$

$$\Theta_{a\phi} = \frac{6a^2((3H^2 + 2k/a^2)/8\pi - V)}{((3/4\pi)(H^2 + k/a^2) - 2V)^{1/2}}. \tag{26}$$

This form is not symplectic (it is degenerate), so one cannot construct a natural volume measure on C. Ideally, the "real" phase space of our system would be given by "solving the dynamics," and then taking equivalence classes of phase points that are part of the same trajectory. In this way one would obtain the space of motions, onto which one could then pull back the degenerate form to obtain a new symplectic form (of degree two less than ω) and construct a canonical measure. This is quite complicated in general due to the differential equation that must be solved. The usual approach to take instead is to set H to some value H_* in the differential form and define their measure accordingly, i.e. set

$$\mathrm{d}\Omega = \omega_{a,H,\phi}|_{H=H_*} = \Theta_{a\phi}|_{H=H_*}\mathrm{d}a\mathrm{d}\phi. \tag{27}$$

One may do this because surfaces of constant Hubble parameter in phase space are transverse to temporal evolution, and the measure is preserved under translation of these surfaces along the Hamiltonian flow. Finally, one may naturally define the GHS measure μ_{GHS} on Lebesgue measurable sets U by

$$U \mapsto \int_U \mathrm{d}\Omega = -6\int_U a^2\frac{(3H_*^2 + 2k/a^2)/8\pi - V}{((3/4\pi)(H_*^2 + k/a^2) - 2V)^{1/2}}\mathrm{d}a\mathrm{d}\phi. \tag{28}$$

This expression of the GHS measure is equivalent to those derived in (Carroll and Tam, 2010; Schiffrin and Wald, 2012).[4]

[4]There are some complications with the $k = 1$ case. See (Schiffrin and Wald, 2012, 8) for the details. I have however chosen not to set $8\pi G = 1$, but rather maintained consistency with the rest of this dissertation's use of "geometrical units" by only setting $G = 1$. Gibbons et al. (1987) use a simplifying, but less transparent coordinate choice. They also choose to investigate only the special case where $V = m^2\phi^2/2$. It can be shown with some work that their expression is equivalent to this one as well with this potential.

2 The Flatness Problem

The GHS measure clearly diverges for large scale factors, a point originally recognized by Gibbons et al. (1987, 745); it also converges to 0 for small scale factors. Due to the divergence, one may readily say that, given any choice of Hubble parameter H_*, almost all spacetimes will have a "large" scale factor. More precisely, pick any scale factor a_*; the set of spacetimes with $a < a_*$ is a negligible set: the total measure of this set is finite whereas the total measure of its complement is infinite. What is the significance of this fact about the GHS measure, specifically for the flatness problem?

Hawking and Page (1988, 803-4) suggest the following:

> "Thus for arbitrarily large expansions (and long times), and for arbitrarily low values of the energy density, the canonical measure implies that almost all solutions of the Friedmann-Robertson-Walker scalar equations have negligible spatial curvature and hence behave as $k = 0$ models. In this way a uniform probability distribution in the canonical measure would explain the flatness problem of cosmology..."

By "arbitrarily large expansions" (and "arbitrarily low values of energy density"), they appear to mean the following. Pick any arbitrary a_* (and any arbitrary ϕ_*).[5] According to the GHS measure almost all spacetimes have $a > a_*$ (and $\phi > \phi_*$), or, equivalently, the spacetimes with $a < a_*$ (and $\phi < \phi_*$) compose a negligible set. Furthermore, since this holds for any choice of a_*, one may infer that almost all spacetimes are arbitrarily close to having $\kappa = 0$ (since $\kappa = k/a^2$) in exactly the same sense. It is perhaps somewhat misleading to say that curved FRW spacetimes with large scale factors "behave as $k = 0$ models;" the curvature does not change in such models. It is, however, surely false to say that a "uniform probability distribution" with respect to the GHS measure would explain the flatness problem of cosmology. There is in fact no such uniform probability distribution, since the GHS measure is not finite. Moroever, there is also no canonical probability distribution ρ at all which would make $U \mapsto \int_U \rho \mathrm{d}\Omega_{GHS}$ into a probability measure—one has to make a choice in order to obtain a probability measure in the case of infinite total measure, a choice which appears completely arbitrary in this context.

Carroll and Tam (2010, 14) invite us to consider the question in more "physically transparent" terms by looking at the curvature κ, which I previously exchanged in favor of the scale factor a when deriving the GHS measure. One can recast the scale factor a as the curvature κ using the relation from before, namely $\kappa = k/a^2$. (Note especially that this switch maps the entire set of scale factors for the $k = 0$ case to the single point $\kappa = 0$.) One then defines the GHS measure (at least for curved FRW spacetimes) by the map

$$U \mapsto \int_U \mathrm{d}\Omega = -6 \int_U \frac{1}{|\kappa|^{5/2}} \frac{(3H_*^2 + 2\kappa)/8\pi - V}{((3/4\pi)(H_*^2 + \kappa) - 2V)^{1/2}} \mathrm{d}\kappa \mathrm{d}\phi. \tag{29}$$

It is clear that the measure diverges for small values of curvature, i.e. curvatures close to flat, due to the curvature term in the denominator. This is pointed out by Carroll and Tam (2010, 15). They suggest the following interpretation of this fact:

> "Considering first the measure on purely Robertson-Walker cosmologies (without perturbations) as a function of spatial curvature, there is a divergence at zero curvature. In other words, curved [FRW] cosmologies are a set of measure zero—the flatness problem, as conventionally understood, does not exist."

[5] Gibbons and Turok (2008, 6) point out that ϕ is always bounded given H_*, so it is not really necessary to pick an arbitrary ϕ_*.

As stated these claims are highly suspect.

Firstly, Carroll and Tam assert that all values of their curvature coordinate Ω_k (essentially equivalent to κ) can be integrated over. While this is perhaps true, portraying the phase space in terms of curvature is misleading. For curved FRW spacetimes, it is true that the measure diverges for small values of curvature κ, as I indicate above and as Hawking and Page suggest in the passage from their paper quoted above. The recast measure, however, is infinite *at* zero curvature because the entire set of $k = 0$ scale factors is mapped to $\kappa = 0$. The GHS measure diverges for large scale factors in the case of flat FRW spacetimes just as it does for curved FRW spacetime. Thus it is misleading to describe a "divergence at zero curvature;" there is nothing special going on in flat FRW spacetimes (at least in this respect).[6]

Secondly (and relatedly), curved FRW spacetimes are clearly not a set of measure zero—at least according to the GHS measure. The initial data of FRW spacetimes is representable in the space $\{a, \dot{a}, \phi, \dot{\phi}, k\}$. The curvature constant k serves as an index for *three different phase spaces*, each of which has an infinite total measure—even after taking into account constraints and choosing a hypersurface in the constraint surface according to GHS's procedure. The unboundedness of the total phase space measure for each kind of FRW spacetime is due, again, to the unbounded range of the scale factor Schiffrin and Wald (2012, 11).[7] This is quite plain when one expresses the GHS measure in terms of the scale factor. Transforming to the curvature coordinate κ should not change the fact that the total measure of each phase space is infinite. So, while it is true that the GHS measure attributes infinite measure to flat FRW spacetimes (as Carroll and Tam appear to recognize), it also does so both to positively curved FRW spacetimes and to negatively curved spacetimes. Therefore it is false that the curved FRW cosmologies are a set of measure zero according to the GHS measure; hence one cannot conclude on this basis that the flatness problem does not exist.

One might try to rescue Carroll and Tam's claim about the flatness problem by interpreting flatness more broadly, namely by including "nearly flat" curved spacetimes. This requires specifying what the set of "nearly flat" curved spacetimes is to be, e.g. a specification of the set of spacetimes with curvature less than some κ_* (at some time corresponding to Hubble parameter H_*). Almost all spacetimes will have a "small" curvature κ in comparison to this curvature κ_*. In other words, the set of spacetimes with $\kappa > \kappa_*$ is a negligible set. Since our universe's spatial curvature is thought to be "nearly flat," i.e. it should be less than κ_* (whatever it is), it follows from this argument that our universe is actually typical, contra what is assumed in the flatness problem. Unfortunately this argument does not follow from the GHS measure alone, since one had to make an independent choice in choosing κ_*, a choice that is not natural in any clear sense whatever. Furthermore, it is doubtful that there is any reasonable argument to justify a choice of κ_*—an explication of "close to flat" in the context of FRW models; it appears to be a completely arbitrary choice.

Here is a slightly different tack into the same stiff headwind. Suppose κ_* is the (non-zero) spatial curvature of our universe at the present time. The GHS measure can be used to infer that almost all spacetimes with the same Hubble parameter will have flatter spatial curvatures. In such circumstances, one might be inclined to wonder "Why is my universe's spatial curvature so large? It seems like it ought to be much smaller if my universe is typical!" On this line of thought, it seems like one actually has a curvature problem rather than a flatness problem. Of course one would say this for any κ_* whatsoever, regardless of its magnitude,

[6] Carroll and Tam appear to equivocate several times between there being a divergence *at* $\kappa = 0$ and the measure diverging *as* $\kappa \to 0$: "The integral diverges near [$\kappa = 0$], which is certainly a physically allowed region of parameter space" (Carroll and Tam, 2010, 17); "The measure diverges on flat universes" (Carroll and Tam, 2010, 28).

[7] Besides in (Schiffrin and Wald, 2012), this fact is correctly pointed out in (Gibbons et al., 1987; Hawking and Page, 1988). While Carroll and Tam (2010, 20-1) observe that "this divergence was noted in the original GHS paper, where it was attributed to 'universes with very large scale factors' due to a different choice of variables," they object to this as an interpretation: "This is not the most physically transparent characterization, as any open universe will eventually have a large scale factor." For this reason they exchange the scale factor for curvature; it is not clear, however, how this characterization is more physically transparent since it amounts to the same thing.

so it is not clear how one would ever be in the position to be satisified with one's curvature in an FRW universe—at least insofar as one expects things in our universe to be typical (in accord with Copernican principle-style reasoning). No matter. The measure suggests this question. What is the answer?

The answer is that the curvature depends on the actual dynamical history of the universe, and so it has no explanation within the context of the HBB model (apart from one depending on an initial condition). That answer may be unsatisfying, but the question is a bad one anyway, driven by misleading intuitions. There is no such thing as a typical FRW spacetime, and the GHS measure is not going to explain why the universe's curvature is what it is. This kind of thinking is clearly motivated by supposing that the GHS measure can be used as a likelihood measure, as Carroll and Tam clearly do:

> "When we consider questions of fine-tuning, however, we are comparing the real world to what we think a randomly-chosen history of the universe would be like" (Carroll and Tam, 2010, 11).

Some popular, specious conceptions (in physics and beyond) of statistical mechanics encourage this line of thought. Putatively successful typicality arguments in statistical mechanics (Goldstein, 2012) depend, however, not only on having a phase space measure, but also on both the dynamics of the system and on a specification of macroproperties or macrostates (defined as regions of phase space) (Frigg, 2009; Frigg and Werndl, 2012). Accordingly, any claim of fine-tuning in FRW spacetimes on the sole basis of the GHS measure (which does at least incorporate the FRW dynamics) is bound to miss the mark without additional assumptions (such as a well-motivated standard of flatness).

Gibbons and Turok (2008) take a different approach from Carroll and Tam. They correctly observe that universes with large scale factors are universes with small spatial curvatures. They then claim that the scale factor is neither "geometrically meaningful" nor "physically observable" and therefore propose to identify all the "indistinguishable" nearly flat spacetimes on the surface identified by H_*.[8] They do so by effectively choosing a "cutoff" curvature κ_* and throwing out all the spacetimes with curvatures smaller than it. The advantage to doing this is that the total measure of FRW spacetimes with curvatures larger than κ_* is finite, so that one can then define a probability measure in a natural way.

The disadvantage is that this makes no sense. Carroll and Tam (2010, 20) comment, "to us, this seems to be throwing away almost all the solutions, and keeping a set of measure zero. It is true that universes with almost identical values of the curvature parameter will be physically indistinguishable, but that doesn't affect the fact that almost all universes have this property." Indeed, doing what Gibbons and Turok do is throwing away almost all the solutions (although the remaining set has finite measure, not measure zero as Carroll and Tam claim). They are also right to point out that if nearly flat universes are physically indistinguishable, so are "nearly-κ" universes for almost any κ. Gibbons and Turok do not throw out these universes however (else they would not have been left with any universes at all). Their justification for an additional assumption therefore fails.

Ironically, Carroll and Tam make essentially the same error as Gibbons and Turok, by identifying the flat and nearly flat spacetimes. Instead of throwing out all the flat and nearly flat spacetimes like the latter pair, however, the former pair throws out the complement of the flat and nearly flat spacetimes by assigning them zero measure. They then triumphantly conclude that all FRW spacetimes are essentially flat! Carroll and Tam propose to tame the remaining divergence in the GHS measure by regularizing the integral, in effect making the measure finite. The problem with doing this is that, since the GHS measure is not finite,

[8] It is not clear what they mean by "geometrically meaningful." The scale factor is clearly geometric in the relevant sense, since it relates spaceforms of the same kind by scalings. It is moreover physically meaningful because space is expanding (or contracting) in FRW spacetimes. The precise value of a does not matter, as it can be re-scaled, but that does not undermine its meaningfulness. It is also unclear how the fact that a is physically unobservable should matter, since most features of spacetime are not observable, e.g. the metric g, the spatial curvature κ, etc. The physically relevant content of these, including the scale factor, can be inferred from observations and appropriate assumptions.

regularizing the measure makes it no longer the GHS measure, in which case any justification the measure had by its "naturalness" is lost since a choice was made.[9] In short, one may as well have just assumed the probability distribution they end up with from the very beginning.Their stated justification for this move is pragmatic: "This non-normalizability is problematic if we would like to interpret the measure as determining the relative fraction of universes with different physical properties" (Carroll and Tam, 2010, 17). However this is obviously an inadequate justification for the propriety of their measure.

References

Albrecht, Andreas, and Paul Steinhardt. "Cosmology for Grand Unified Theories with Radiatively Induced Symmetry Breaking." *Physical Review Letters* 48: (1982) 1220–1223.

Belinsky, Vladimir, Leonid Grishchuk, Isaak Khalatnikov, and Yakov Zeldovich. "Inflationary Stages in Cosmological Models with a Scalar Field." *Physics Letters B* 155: (1985) 232–236.

Carroll, Sean. "In What Sense Is the Early Universe Fine-Tuned?" In *Time's Arrows and the Probability Structure of the World*, edited by Barry Loewer, Brad Weslake, and Eric Winsberg, Cambridge, MA: Harvard University Press, forthcoming.

Carroll, Sean, and Heywood Tam. "Unitary Evolution and Cosmological Fine-Tuning." ArXiv Eprint, 2010. `http://arxiv.org/abs/1007.1417`.

Coule, David. "Canonical measure and the flatness of a FRW universe." *Classical and Quantum Gravity* 12: (1995) 455–469.

Dicke, Robert, and Jim Peebles. "The Big Bang Cosmology—Enigmas and Nostrums." In *General Relativity: An Einstein Centenary Survey*, edited by Stephen Hawking, and Werner Israel, Cambridge: Cambridge University Press, 1979, chapter 9, 504–517.

Frigg, Roman. "Typicality and the Approach to Equilibrium in Boltzmannian Statistical Mechanics." *Philosophy of Science* 76: (2009) 997–1008.

Frigg, Roman, and Charlotte Werndl. "Demystifying Typicality." *Philosophy of Science* 79: (2012) 917–929.

Geroch, Robert. "General Relativity.", 1972. Unpublished lecture notes.

Gibbons, Gary, Stephen Hawking, and John Stewart. "A natural measure on the Set of all Universes." *Nuclear Physics B* 281: (1987) 736–751.

Gibbons, Gary, and Neil Turok. "Measure problem in cosmology." *Physical Review D* 77: (2008) 1–12.

Goldstein, Sheldon. "Typicality and Notions of Probability in Physics." In *Probability in Physics*, edited by Yemima Ben-Menahem, and Meir Hemmo, Berlin: Springer Verlag, 2012, chapter 4, 59–71.

Guth, Alan. "Inflationary universe: A possible solution to the horizon and flatness problems." *Physical Review D* 23, 2: (1981) 347–356.

[9]Carroll more recently has conceded the artificiality of regularizing: "Earlier attempts to regularize the measure, for example by considering an ϵ-neighborhood around the zero-curvature Hamiltonian constraint surface (Carroll and Tam, 2010) or by identifying universes with similar curvatures (Gibbons and Turok, 2008) have not proven satisfactory" (Remmen and Carroll, 2013, 7). He remains convinced, however, that almost all FRW spacetimes are "nearly flat:" "we should throw all of the others away and deal with flat universes," (Carroll, forthcoming, 19), developing a measure on just these spacetimes in a later paper (Remmen and Carroll, 2014).

Hawking, Stephen, and Don Page. "How Probable is Inflation?" *Nuclear Physics B* 298: (1988) 789–809.

Linde, Andrei. "A New Inflationary Universe Scenario: A Possible Solution of the Horizon, Flatness, Homogeneity, Isotropy, and Primordial Monopole Problems." *Physics Letters B* 108: (1982) 389–393.

———. "The inflationary universe." *Reports on Progress in Phyics* 47: (1984) 925–986.

Malament, David. *Topics in the Foundations of General Relativity and Newtonian Gravity Theory*. Chicago: University of Chicago Press, 2012.

McCabe, Gordon. "The structure and interpretation of cosmology: Part I—general relativistic cosmology." *Studies in History and Philosophy of Modern Physics* 35: (2004) 549–595.

McCoy, Casey. "Does inflation solve the hot big bang model's fine-tuning problems?" *Studies in History and Philosophy of Modern Physics* 51: (2015) 23–36.

Remmen, Grant, and Sean Carroll. "Attractor solutions in scalar-field cosmology." *Physical Review D* 88: (2013) 1–14.

———. "How many *e*-folds should we expect from high-scale inflation?" *Physical Review D* 90: (2014) 1–14.

Schiffrin, Joshua, and Robert Wald. "Measure and probability in cosmology." *Physical Review D* 86: (2012) 1–20.

Smeenk, Chris. "Philosophy of Cosmology." In *The Oxford Handbook of Philosophy of Physics*, edited by Robert Batterman, Oxford: Oxford University Press, 2013, chapter 17, 607–652.

Wald, Robert. *General Relativity*. Chicago: University of Chicago Press, 1984.

Wolf, Joseph. *Spaces of Constant Curvature*. Providence, RI: AMS Chelsea Publishing, 2010, 6th edition.

Invariance, Interpretation, and Motivation

Thomas Møller-Nielsen

July 2016

[Forthcoming in *Philosophy of Science (2016 Proceedings)*.]

Abstract

In this paper I assess the 'Invariance Principle', which states that only quantities that are invariant under the symmetries of our theories are physically real. I argue, contrary to current orthodoxy, that the variance of a quantity under a theory's symmetries is not a sufficient basis for interpreting that theory as being uncommitted to the reality of that quantity. Rather, I argue, the variance of a quantity under symmetries only ever serves as a motivation to refrain from any commitment to the quantity in question. In the process of this discussion, I address the related but importantly distinct issue of when symmetries can be said to prompt a mathematical reformulation of the relevant theory.

1 Introduction

Take the *Invariance Principle* to be the principle that only quantities that are invariant under the symmetries of our theories are physically real.[1] It is a doctrine with a distinguished pedigree: acclaimed theorists as diverse as the physicist Paul Dirac, the mathematician Hermann Weyl, and the philosopher Robert Nozick were all apparent signatories during their respective lifetimes.[2] *Prima facie*, however, it is something of a mystery as to how and why the principle is supposed to work. Nevertheless, there appear to be at least some uncontroversial cases where it—or something very close to it—does work.

One such example can be found in Newtonian Gravitation Theory (NGT), i.e., the theory comprising Newton's three laws, plus his inverse square gravitational law, governing the behaviour of point particles in Newtonian spacetime. As is well known, this theory is *Galilean invariant*. This implies, among other things, that if one takes any solution to NGT and "boosts" it—that is, uniformly alters the absolute velocity of each point particle by the same amount throughout its history—one will invariably get back a solution to NGT. Boosts, in other words, are a *symmetry* of NGT: they are transformations that invariably map solutions of the theory to solutions.

[1] I draw the term from Saunders (2007). Compare also Dasgupta's (forthcoming) "symmetry-to-reality inference".

[2] See, e.g., Dirac (1930, vii), Weyl (1952, 132), and Nozick (2001, 82).

Which quantity varies under this particular symmetry? The answer is obvious: absolute velocity. Thus, according to the Invariance Principle, we should conclude that absolute velocity is not a genuine physical quantity. Conversely, which quantities are invariant under this particular symmetry? Again, the answer is obvious: relative (inter-particle) distance and velocity, temporal intervals, and absolute acceleration. Thus, according to the Invariance Principle, we should conclude that NGT's boost symmetry does not threaten these quantities' status as genuinely physical.

As it turns out, one can successfully purge Newtonian theory of the spacetime structure required to make absolute velocity a physically meaningful quantity. More specifically, one can move to *Galilean spacetime*. (Sometimes also called "Neo-Newtonian spacetime".)[3] Here, the Newtonian posit of persisting points of absolute space—persisting points which, crucially, allow for the notion of absolute velocity to be physically meaningful—is done away with, but an *affine structure* is nevertheless preserved, which defines the "straight" or force-free (inertial) paths through spacetime. Absolute velocity is therefore not a physically meaningful quantity in Galilean spacetime, as it is in Newtonian spacetime. Nevertheless, all other Newtonian notions, including the notion of absolute acceleration, remain well-defined in Galilean spacetime. To the extent that one opts for Galilean over Newtonian spacetime, then, one has excised an ostensibly odious piece of theoretical structure from NGT.

Three important caveats are worth noting, however. First, and most obviously, none of this is to say that Newtonian theory set in Galilean spacetime is therefore the true and complete theory of the world. (It isn't.) Second, nor is this to say that by moving to Galilean spacetime one has thereby purged Newtonian theory of all its "variant" structure. (One hasn't. The symmetry group of Newtonian theory is actually wider than the Galilean group: it has additional symmetries.)[4] Third, nor is this even to say that the invariant quantities one ends up with following such an application of the Invariance Principle will invariably be preserved in future theories. (For instance, there is no notion of "relative spatial distance" *simpliciter* in special relativity.) Given all of these caveats, however, one might well ask: What good is the Invariance Principle, exactly? What purpose, in particular, does it serve?

As I see it—and, I take it, as many other contemporary theorists also see it—the purpose of the Invariance Principle is essentially *comparative*. That is, it is simply supposed to lead you to a *better theory*—or a better interpretation, or characterisation, of the same theory—than the one you started with. To take the case at hand: Newtonian theory set in Galilean spacetime is a better theory than Newtonian theory set in Newtonian spacetime. It is a theory which possesses all of the theoretical virtues of its rival, but lacks any apparent ontological commitment to the unwanted variant quantity in question.

In summary, the Galilean invariance of NGT, in conjunction with the Invariance Principle, is supposed to indicate that neither absolute velocity nor

[3] See, e.g., Earman (1989, §2.4).

[4] See, e.g., Knox (2014). I discuss this point further in Section 4 below.

any corresponding persisting points of absolute space are genuinely real. Now to lay my cards on the table: I actually think that something *very close* to this general kind of inference—that is, from the variance of a quantity under symmetries to that quantity's nonreality—is legitimate. The devil, however, is in the details. In particular, I don't believe that the *mere* Galilean invariance of NGT is enough to establish absolute velocity's nonreality. And in general, I don't believe that the *mere* variance of a quantity under symmetries is enough to establish that quantity's nonreality. These beliefs, as far as I can determine, put me in the minority camp in the contemporary philosophical literature on symmetries. Nevertheless, I think they are correct beliefs—and they are precisely the ones that I will attempt to argue for in the remainder of this paper.

2 Interpretational vs Motivational

In arguing for the above claims, it will prove extremely useful first to distinguish between two very different ways of thinking about symmetries.

Close cousins of the distinction that I have in mind have already been drawn in the literature. Thus, Greaves and Wallace write:

> There is a widespread consensus that two states of affairs related by a symmetry transformation are really just the same state of affairs differently described. That is, if two mathematical models of a physical theory are related by a symmetry transformation, then those models represent one and the same physical state of affairs. (Greaves and Wallace 2014, 60)

They continue:

> Although we agree with this consensus [...] even those who do not agree that symmetry-related states of affairs are identical at least agree that they are *empirically indistinguishable* from one another. (Greaves and Wallace 2014, 60, fn 1)

To illustrate the difference between these two ways of thinking about symmetries, consider again the example of boosts in NGT. According to the "widespread consensus" view alluded to, and endorsed by, Greaves and Wallace, boosted models of NGT are to be taken to represent the same physical state of affairs *even when the theory is putatively set in Newtonian spacetime*. In other words, according to this view, one needn't make the move to Galilean spacetime in order not to be committed to absolute velocities; there is a way of understanding boosted models' physical equivalence, and their associated noncommitment to the notion of absolute velocity, prior to making this move.[5]

Things are very different according to the second conception of symmetries described, and rejected, by Greaves and Wallace. According to this view, boosted models of NGT are to be regarded as physically *in*equivalent: they are not to be construed as representing the same physical state of affairs. Instead,

[5]See, e.g., Healey (2007, 114-7), for an endorsement of this view in the Newtonian context.

such models are taken to represent physically distinct scenarios, which differ in what absolute velocity they ascribe to the world's total material content. Nevertheless, such models still represent *empirically indistinguishable* states of affairs: in a Newtonian universe, no experiment could ever help an observer determine what her absolute velocity actually is. Such boosted models therefore represent physically distinct ways for the world to be, albeit ones that are indiscernible on the basis of measurement.[6]

As previously mentioned, this distinction between different ways of thinking about symmetries is close, but not identical, to the one that I want to draw. The key reason why it is not identical is because Greaves and Wallace say nothing to the effect that the person who subscribes to the second conception of symmetries—that is, who believes that symmetry-related models invariably represent empirically indistinguishable, but not necessarily physically equivalent, states of affairs—should still be *motivated to seek* an alternative theory, or an alternative interpretation or characterisation of the same theory, according to which such models do not merely represent empirically indistinguishable scenarios, but rather represent physically equivalent states of affairs.[7] Moreover, I claim, it is precisely this notion of *motivation* which plays a central role in correctly understanding the philosophical significance of symmetries in the general case.[8]

Here, then, is what I take to be the appropriate distinction between these two different ways of thinking about symmetries:

- **Interpretational**: Symmetries allow us to *interpret* theories as being committed solely to the existence of invariant quantities, even in the absence of a metaphysically perspicuous characterisation of the reality which is alleged to underlie symmetry-related models.
- **Motivational**: Symmetries only *motivate* us to find a metaphysically perspicuous characterisation of the reality which is alleged to underlie symmetry-related models, but they do not allow us to interpret that theory as being solely committed to the existence of invariant quantities in the absence of any such characterisation.

The central claim of this paper may now be neatly summarised: the (orthodox) interpretational view is mistaken; the (unorthodox) motivational view is correct.

Drawing the distinction in the way that I have done, however, invites the rather obvious question: What, precisely, is meant by a "metaphysically perspicuous characterisation" of reality? This is the question addressed in the next section.

[6] See, e.g., Maudlin (1993, 192), for an endorsement of this view in the Newtonian context.

[7] Compare (again) Maudlin's (1993, 192) discussion in the Newtonian context.

[8] Note that I do not intend any of this as a criticism of Greaves and Wallace's paper. Indeed, as Greaves and Wallace (2014, 60, fn 1) are careful to remark, the distinction they draw is orthogonal to the central topic of their paper, namely the issue of which symmetries have "direct empirical significance" (i.e., have analogues to Galileo's ship).

3 More on Metaphysical Perspicuity

In intuitive terms, a metaphysically perspicuous characterisation of reality is one which corresponds to, or "limns", reality's structure in some suitably faithful way. To use another common (Platonic) metaphor, a metaphysically perspicuous characterisation of reality is one which "carves nature at its joints". (In comparative terms: a description of reality is *more* metaphysically perspicuous than another precisely to the extent that it corresponds to, or limns, reality's structure *more* faithfully than its rival does.)

As many readers will be aware, such a notion is frequently alluded to, and made use of, in contemporary analytic metaphysics.[9] But metaphysical perspicuity is also, I think, a notion that is reasonably serviceable in physical (rather than "merely metaphysical") contexts. One particularly illustrative example—albeit a slightly misleading one, for reasons that I will soon explain—drawn from physics may plausibly be found in classical electromagnetism.[10] As is well known, this theory may be formulated in two different ways.[11] According to one such formulation, EM_1, the theory is expressed in terms of the Faraday tensor, F_{ab}, satisfying the (Maxwell) equations $\nabla_{[a}F_{bc]} = 0$ and $\nabla_a F^{ab} = J^a$, where J^a is a vector field representing the charge current density. According to the second formulation, EM_2, however, the theory is expressed in terms of the vector potential, A_a, satisfying the equation $\nabla_a\nabla^a A^b - \nabla^b\nabla_a A^a = J^b$.

These two formulations of electromagnetism are related to one another. In particular, any model $\langle M, \eta_{ab}, A_a\rangle$ of EM_2 corresponds to a unique model $\langle M, \eta_{ab}, F_{ab}\rangle$ of EM_1, via the equation $F_{ab} = \nabla_{[a}A_{b]}$. The converse, however, is not true. That is, a typical model of EM_1 does *not* typically correspond to a unique model of EM_2. More specifically, if $\langle M, \eta_{ab}, A_a\rangle$ is a model of EM_2 corresponding to a model $\langle M, \eta_{ab}, F_{ab}\rangle$ of EM_1, then so will any other model of EM_2 $\langle M, \eta_{ab}, A'_a\rangle$, where A'_a is related to A_a by a "gauge transformation" $A'_a = A_a + \nabla_a\chi$, where χ is some smooth scalar field.

It is EM_1 which, I take it, constitutes the metaphysically perspicuous characterisation of this theory. That is, it is the tensor F_{ab} which faithfully represents the fundamental ontology of the theory, namely the electromagnetic field. Not so EM_2. This second formulation may, of course, be useful for various calculational or heuristic purposes. But the key point is that the vector potential A_a *does not directly represent a genuinely real field*: rather, it is merely a mathematically convenient "shorthand" way of characterising and determining the values of the Faraday tensor, which *is* taken to represent the genuine material ontology of the theory.[12] Moreover, it is precisely by construing the vector potential in this

[9] See, e.g., O'Leary-Hawthorne and Cortens (1995, 154-7).

[10] Here and below, I take this theory to be set in Minkowski spacetime. Thus, the spacetime models of this theory are of the form $\langle M, \eta_{ab}\rangle$, where M is a four-dimensional differentiable manifold, and η_{ab} is the Minkowski metric.

[11] For a recent, intriguing study of the relationship between these two different formulations of electromagnetism, see Weatherall (forthcoming). I draw heavily on his discussion over the next couple of paragraphs.

[12] Modulo, that is, certain concerns that arise as a result of the Aharonov-Bohm effect. See, e.g., Healey (2007).

way which plausibly allows us to explain and understand, in a fully transparent way, gauge-symmetry models' physical equivalence in EM_2—namely, for the reason that they are merely notationally distinct ways of representing the same fundamental physical ontology.

As mentioned above, I think this example of metaphysical perspicuity is apt to be slightly misleading, at least when taken on its own. This is because this example might make it seem as though having a metaphysically perspicuous characterisation of the (putative) reality underlying symmetry-related models crucially relies upon one having to *mathematically reformulate* the relevant theory (or at least upon having such a mathematical reformulation already in hand), and in particular upon having to reformulate the theory so as to remove any relevant representational redundancy. However, I think this is incorrect. That is, I believe that one *can*, in fact, be in possession of a metaphysically perspicuous characterisation of the reality underlying symmetry-related models *even in the absence* of any mathematical (re-)formulation of the theory which removes the relevant representational redundancy.

Let me illustrate this point with two simple examples. First, consider the case of *shift symmetry* in NGT. This symmetry is subtly different from the case of boost symmetry, discussed above. Here, instead of uniformly altering the absolute velocity of each particle throughout its history, one enacts a global, time-independent repositioning of all matter in space. Thus, for instance, in the shifted world all of the world's material content will (*prima facie*) be located three metres to the left of where it is in the original world. The basic idea behind the "Leibniz shift" argument—the famous argument associated with this symmetry—is that the substantivalist's admission of points of space as primitive objects (allegedly) has the undesirable consequence of committing her to regarding shifted worlds as physically distinct, yet nevertheless empirically indistinguishable:[13] in intuitive terms, everything would look, feel, taste, touch and sound the same in the two (putatively distinct) shifted worlds, just as in the case of boosted worlds.

It will prove helpful to express all of this in terms of the models of the theory. Thus, take a generic model of NGT to be of the form $\mathcal{M} = \langle M, t_{ab}, h^{ab}, \sigma^a, \rho, \phi \rangle$, where M is a differentiable 4-dimensional manifold, t_{ab} is the temporal metric, h^{ab} is the spatial metric, σ^a is the timelike vector field whose integral curves represent the persisting points of absolute space, and ρ and ϕ represent the matter density and the gravitational potential field respectively.[14] A shift symmetry can then be characterised as the application of the appropriate diffeomorphism (corresponding to a spatial translation) d so as to yield a new model $\mathcal{M}_{static} = \langle M, t_{ab}, h^{ab}, \sigma^a, d^*\rho, d^*\phi \rangle$. It is then alleged that $\mathcal{M}$ and $\mathcal{M}_{static}$ differ precisely

[13] Though see Maudlin (1993), who notes that there is an interesting (epistemological) sense in which shifted worlds in NGT are not indiscernible after all.

[14] Note that the canonical presentations of Newtonian spacetime (e.g., Earman 1989, §2.5) take the affine connection as ideologically primitive. I find such presentations unsatisfactory for historical rather than for philosophical reasons: in particular, it threatens to make the move to Galilean spacetime seem almost trivial, and the associated timelike vector field trivially superfluous. For more on this point, see Pooley (MS, §4.4-§4.5).

insofar as they each represent the world's matter content as being located at distinct places in absolute space. More specifically, such Leibniz-shifted scenarios are alleged to differ precisely with regard to which particular points of space are underlying various parts of the matter fields.

For a second example, consider *diffeomorphism symmetry* in general relativity (GR). Here, similarly, the existence of this symmetry is alleged to commit the substantivalist to a plurality of physically distinct possibilities that are nevertheless empirically indistinguishable. In terms of the models of the theory: taking a generic model of GR to be of the form $\mathcal{M} = \langle M, g_{ab}, T_{ab} \rangle$ and applying an arbitrary diffeomorphism d to yield a new model $\mathcal{M}_{diff} = \langle M, d^*g_{ab}, d^*T_{ab} \rangle$ (where M is again a differentiable 4-dimensional manifold, g_{ab} is the metric tensor, and T_{ab} is the stress-energy tensor which, roughly speaking, represents the model's matter content), the two scenarios represented are alleged to differ with regard to which particular points of the spacetime manifold are underlying various parts of the metric and matter fields.[15]

It is my contention that neither the shift symmetry of NGT, nor the diffeomorphism symmetry of general relativity, by themselves motivate any mathematical reconstrual of the respective theories. This is because I believe there is a perfectly transparent, anti-haecceitist, "modestly structuralist"—but nevertheless fully substantivalist—way of understanding such models' representational equivalence even in the absence of any such mathematical reformulation. On this view, spacetime points are construed as genuinely real, fundamental entities. However, they are "contextually individuated": they are not to be understood as being anything more—or less—than "nodes" in the relational, geometrical structures in which they are embedded. Shifted models in NGT and diffeomorphically-related models in GR are thus to be understood as representing the same physical state of affairs precisely because the exact same pattern of relational, geometrical structures is represented as obtaining in each case. Moreover, this view denies that there are any primitive, singular ("haecceitistic") facts about spacetime points which would even allow for a distinction between shifted or diffeomorphically-related scenarios to be coherently drawn.[16]

Whence the difference, then, between the case of gauge symmetry in electromagnetism on the one hand, and shift and diffeomorphism symmetry in NGT and GR on the other? I think the answer is straightforward. In the latter cases, the models in question are *isomorphic*: they represent worlds which differ at most with regard to which particular objects are playing which qualitative roles, i.e., they represent at most haecceitistically distinct possible worlds. Hence, adopting modest structuralism (which implies anti-haecceitism) about spacetime transparently collapses the number of possibilities represented by these models to one. In the former such case, however, the relevant models are *not* isomorphic—read "literally", gauge-related models of EM_2 assign *qualitatively distinct* arrangements of the vector field over spacetime—hence adopting a modestly structuralist ontology does not by itself collapse the number of represented

[15]For further details see, e.g., Earman (1989, §9).

[16]For further defence of this view—which is sometimes also called *sophisticated substantivalism* in the literature—see, e.g., Saunders (2003), Ladyman (2007), and Pooley (2013).

possibilities to one. In order to transparently understand such models' physical equivalence, then, a mathematical reformulation of the theory is required.

To summarise the claims made thus far: according to the motivational view of symmetries, one is invariably only motivated to regard symmetry-related models as physically equivalent; moreover, one is justified in regarding such models as physically equivalent only insofar as one is in possession of a metaphysically perspicuous characterisation of the reality which is alleged to underlie them. However, it is possible to be in possession of a metaphysically perspicuous characterisation of the reality underlying symmetry-related models even in the absence of a mathematical formulation of the theory which removes the relevant representational redundancy. Such a metaphysically perspicuous characterisation is possible just in case the symmetry-related models in question are isomorphic, or are naturally understood as representing at most haecceitistically distinct possibilities. In brief: symmetry-related, isomorphic models invariably do *not* motivate a mathematical reformulation of the relevant theory (modest structuralism invariably suffices); but symmetry-related, *non*-isomorphic models invariably *do*.[17]

4 In Defence of the Motivational View

Let us return once more to the case of NGT. As alluded to in Section 1, the symmetry group of this theory is quite large. For not only does it include transformations corresponding to global velocity boosts of solutions' matter content, but it also includes transformations corresponding to time-dependent translational accelerations of such content (so long as the gravitational potential field is also appropriately transformed). Thus, read "literally", the symmetries of this theory include transformations that map solutions to solutions that represent physically distinct, but nevertheless empirically indistinguishable, states of affairs in which a given material system is:

1. Force-free and stationary with respect to absolute space.
2. Force-free and moving at constant absolute velocity.
3. Absolutely accelerating under a gravitational force-field.

According to the interpretational conception of symmetries, we may legitimately take all of these symmetry-related solutions to in fact represent the same physical state of affairs—despite the fact that they are naturally understood as representing radically distinct physical situations. Things are very different, however, according to the motivational conception of symmetries. On this view, we are merely *motivated to regard* all such solutions as representing the same physical state of affairs, the motivation arising from the general Occamist principle that, other things being equal, our preferred scientific theories should not allow for solutions that represent physically distinct but nevertheless empirically indistinguishable possible worlds. According to the motivational

[17] See also Pooley (2013, 576-7) and Weatherall (forthcoming) for recent, related arguments to this effect.

view, then (and to repeat slightly), absent a metaphysically perspicuous characterisation of the reality underlying these symmetry-related models, we have no choice but to regard them as representing physically distinct states of affairs.

For our purposes, the crucial thing to note about all of these models is that *none of them are isomorphic*—naturally understood, they do not represent at most haecceitistically distinct possible worlds. According to the criterion laid down in the previous section, then, in order to be able to transparently understand how it could be that such models may be said to represent physically equivalent scenarios, a mathematical reformulation of the theory is required.

As it turns out, such a mathematical reformulation of the theory is possible. In brief, in this reformulation one replaces the vector field σ^a with a new kind of *dynamical* inertial connection ∇^{NC}, with models of the form $\mathcal{M}_{NC} = \langle M, t_{ab}, h^{ab}, \nabla^{NC}, \rho \rangle$. Up to isomorphism, any two symmetry-related models of NGT correspond to a unique model of Newtonian gravity geometrised in this way. Thus, it is said, by moving to this "Newton-Cartan" theory one successfully removes the undesirable "gauge-redundancy" inherent in all non-geometrised versions of Newtonian gravitation theory.[18]

What might the defender of the interpretational view of symmetries say in defence of her view—in this context, that the move to Newton-Cartan theory is not required in order to be able to legitimately regard all symmetry-related solutions of NGT as physically equivalent?

I anticipate two likely lines of response. First, she might attempt to establish the preferability of her view over the motivational view by noting that the defender of the motivational view is committed, at least prior to the appropriate theory's reformulation (in the context of NGT), to the existence of in principle undetectable (symmetry-variant) matters of fact. Moreover, the defender of the interpretational view might argue, this is an unpalatable consequence, one which we would do best to avoid—and one which, she might point out, the interpretational view does in fact avoid.

I agree that the admission of such in principle undetectable facts is an undesirable consequence of the motivational view. However, I do not think that this admission is sufficiently unpalatable so as to be capable of refuting the motivational view, or even of establishing the preferability of the interpretational view over the motivational view. After all, prohibitively strong versions of verificationism aside, there is nothing obviously absurd about admitting in principle undetectable facts into one's ontology; nor is there any obvious reason why we should always be capable of discovering a theory, or a perspicuous characterisation thereof (the case of isomorphic models excepted), which succeeds in transparently explaining such solutions' empirical equivalence by virtue of

[18]For further details, see, e.g., Knox (2014). Note also the important point that moving to Newton-Cartan theory is not by itself sufficient for one to be able to transparently understand as physically equivalent all symmetry-related models of Newtonian theory set in flat spacetime. This is because—as mentioned above—such symmetry-related models will typically correspond to a single model of Newton-Cartan theory *only up to isomorphism*. Thus, in order to have a *fully* transparent understanding of how it is that symmetry-related models of Newtonian theory set in flat spacetime can correspond to a single model of Newton-Cartan theory, a modestly structuralist conception of spacetime ontology is also required.

their actual physical equivalence; nor indeed is there even any obvious way of guaranteeing that there will always be such a theory or characterisation (again, isomorphic models excepted) waiting in logical space to be discovered.

Furthermore, although it is to be admitted that the Newtonian who subscribes to the merely motivational view of symmetries might indeed be committed to the possibility of there being facts beyond her epistemic grasp, it nevertheless bears emphasising that for such a Newtonian there is a perfectly good explanation as to *why* such facts are epistemically inaccessible: they are inaccessible precisely because the world is in fact accurately described by the laws of NGT, with associated models of the form $\langle M, t_{ab}, h^{ab}, \sigma^a, \rho, \phi \rangle$, and because all any Newtonian observer ultimately has empirical access to are the relative distances and velocities between material entities. For such a Newtonian, then, the empirical phenomena underdetermine the genuine physical facts; but the theory itself is able to provide a perfectly transparent explanation of the reality behind the phenomena in terms of which the underdetermination can be straightforwardly understood.

The Newtonian who adopts the interpretational construal of symmetries, however, would appear to lose this explanatory transparency. In other words, she might know *that* she may legitimately regard all symmetry-related solutions as physically equivalent; but the reality in terms of which this physical equivalence is to be understood will (absent a reformulation of the theory) remain opaque to her; she is offered no immediate explanation as to *how* such physical equivalence is to be construed, or how it could even be said to arise.

These considerations naturally suggest a second possible line of response for the defender of the interpretational view. In particular, she might claim that she *does*, in fact, have a transparent understanding of the reality underlying NGT's symmetry-related models, and that such a transparent understanding is in fact attainable *prior* to the move to Newton-Cartan theory.[19]

Such a response evidently leads into deep philosophical waters very quickly. (After all, what does it mean to be in possession of a "transparent understanding" of anything?) But let me make a brief remark as to why I find this particular claim to be implausible. For note that in NGT the persisting points of absolute space are not merely "idly turning wheels" that can simply be expunged from the theory without explanatory loss: they are not "explanatorily idle" posits. This is for two main reasons. First, such points play a crucial role in the *metaphysical* explanation of what quantities like relative velocity and absolute rotation and absolute acceleration truly are: for the Newtonian, facts about particular inter-particle velocities and absolute rotations and absolute accelerations are naturally understood as being *grounded in* particular facts about (rates of change of) absolute velocities.[20] Second, such points provide the crucial transtemporal standard which is required in the realist's *causal* explanation of the observable effects of noninertial motion (e.g., Newton's famous "bucket experiment"): a standard without which Newton's laws simply cannot be formu-

[19] Dewar (2015, esp. 322)—who is a recent, explicit defender of the interpretational view—is plausibly read as making this claim.

[20] Cf. Pooley (MS, 118).

lated (at least, absent any *other* way of construing the transtemporal structure required to underwrite the distinction between inertial and noninertial motion). In short—and to the extent that the interpretational view is not supposed to reduce to a rather uninteresting form of scientific instrumentalism—it is simply not clear what causal-explanatory, *realistic* picture of the world is being propounded by the defender of the interpretational view, at least in this particular (Newtonian) context; it is simply opaque what, according to her, *the world is really like*.

Acknowledgements

For extremely helpful comments and discussion, I would like to thank Neil Dewar, James Ladyman, Niels Martens, Tushar Menon, Oliver Pooley, James Read, Simon Saunders, Alex Skinner, Teru Thomas, David Wallace, and audiences in London and Cardiff.

References

Dasgupta, S. (forthcoming), "Symmetry as an Epistemic Notion (Twice Over)." *British Journal for the Philosophy of Science.*

Dewar, N. (2015), "Symmetries and the Philosophy of Language." *Studies in the History and Philosophy of Modern Science*, Vol. 52, pp. 317-327.

Dirac, P. A. M. (1930), *The Principles of Quantum Mechanics.* Oxford University Press. (Reference is made to 1958 (4th) edition.)

Earman, J. (1989), *World-Enough and Space-Time.* MIT Press.

Greaves, H. and Wallace, D. (2014), "Empirical Consequences of Symmetries." *British Journal for the Philosophy of Science*, Vol. 65, No. 1, pp. 59-89.

Healey, R. (2007), *Gauging What's Real.* Oxford University Press.

Knox, E. (2014), "Newtonian Spacetime Structure In Light of the Equivalence Principle." *British Journal for the Philosophy of Science*, Vol. 65, No. 4, pp. 863-880.

Ladyman, J. (2007), "Scientific Structuralism: On the Identity and Diversity of Objects in a Structure." *Aristotelian Society Supplementary Volume*, Vol. 81, No. 1, pp. 23-43.

Maudlin, T. (1993), "Buckets of Water and Waves of Space: Why Spacetime Is Probably a Substance." *Philosophy of Science*, Vol. 68, No. 2, pp. 183-203.

Nozick, R. (2001), *Invariances: The Structure of the Objective World.* Harvard University Press.

O'Leary-Hawthorne, J. and Cortens, A. (1995), "Towards Ontological Nihilism." *Philosophical Studies*, Vol. 79, No. 2, pp. 143-165.

Pooley, O. (2013), "Substantivalist and Relationalist Approaches to Spacetime." In R. Batterman (ed.), *Oxford Handbook of Philosophy of Physics.* Oxford University Press.

Pooley, O. (MS), *The Reality of Spacetime.* Book manuscript.

Saunders, S. (2003), "Physics and Leibniz's Principles." In K. Brading & E. Castellani (eds.), *Symmetries in Physics: Philosophical Reflections.* Cambridge University Press.

Saunders, S. (2007), "Mirroring as an A Priori Symmetry." *Philosophy of Science*, Vol. 74, No. 4, pp. 452-480.

Weatherall, J. (forthcoming). "Understanding Gauge." *Philosophy of Science.*

Weyl, H. (1952), *Symmetry.* Princeton University Press.

Atlanta, GA; 3-5 November 2016 -407-

Black Holes, Information Loss and the Measurement Problem

Elias Okon
Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México, Mexico City, Mexico.

Daniel Sudarsky
Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Mexico City, Mexico.

The *information loss paradox* is often presented as an unavoidable consequence of well-established physics. However, in order for a genuine paradox to ensue, not-trivial assumptions about, e.g., quantum effects on spacetime, are necessary. In this work we will be explicit about these additional, speculative assumptions required. We will also sketch a map of the available routes to tackle the issue, highlighting the, often overlooked, commitments demanded of each alternative. In particular, we will display the strong link between black holes, the issue of information loss and the measurement problem.

1 Introduction

The so-called *information loss paradox* is usually introduced as an unavoidable consequence of standard, well-established physics. The paradox is supposed to arise from a glaring conflict between Hawking's black hole radiation and the fact that time evolution in quantum mechanics preserves information. However, the truth is that, in order for a genuine paradox to appear, a sizable number of additional, non-standard assumptions is required. As we will see, these extra assumptions involve thesis regarding the fundamental nature of Hawking's radiation, guesses regarding quantum aspects of gravity and even considerations in the foundations of quantum theory.

In this work, we will be explicit about the additional assumptions required for a genuine conflict to arise and delineate the available options in order to tackle the issue. In particular, we will stress the connection between information loss and the measurement problem, and display the often non-trivial commitments that each of the available alternatives to solve the information loss issue demands.

2 The classical setting: black holes hide information

We start by reviewing some properties of classical black holes. Gravity, being always attractive, tends to draw matter together to form clusters. In fact, if the mass of a cluster is big enough, nothing will be able to stop the contraction until, eventually, a black hole will form. That is, the gravitational field at the surface of the body will be so strong that not even light will be able to escape and a region of spacetime from which nothing is able to emerge will form. The boundary of such a region is called the event horizon and, according to general relativity, its area never decreases.

In general, the collapse dynamics that leads to the formation of a black hole can, of course, be very complicated. However, it can be shown that all such systems eventually settle down into one of the few stationary black hole solutions, which are completely characterized by the mass, charge and angular momentum of the the Kerr-Newman spacetimes. In fact, the so-called black hole uniqueness theorems guarantee that, as long as one only considers gravitational and electromagnetic fields, then these solutions represent the complete class of stationary black holes. Moreover, the so-called no-hair theorems ensure that the set of stationary solutions does not grow, even if one considers other hypothetical fields.

The above mentioned results seem to suggest that when a cluster collapses to form a black hole, a large amount of information is lost. That is, details such as the multipole moments of the initial mass distribution, or the type of matter involved, seem to be altogether lost when the black hole settles. Note however that such apparent loss of information corresponds only to that available to observers outside of the black hole. While at early times there are Cauchy hypersurfaces[1] completely contained outside of the black hole, at later times all Cauchy hypersurfaces have parts both inside and outside it (see Figure 1). Therefore, using data located both outside and inside of the black hole, the *whole* spacetime can always be recovered. We conclude that, in the classical setting, information is not really lost. All that happens is that, when a black hole forms, a new region of no escape emerges and some of the information from the outside of the black hole moves into such new region. One could still argue that, since there are points inside of the horizon which are not in the past of future null infinity,[2]

[1] A Cauchy hypersurface is a subset of spacetime which is intersected exactly once by every inextensible, non-spacelike curve.

[2] Future null infinity is the set of points which are approached asymptotically by null rays which

then it is impossible to reconstruct the whole spacetime by evolving backwards the data on it. However, future null infinity is not a Cauchy hypersurface so one should not expect to reconstruct the whole spacetime from such data.

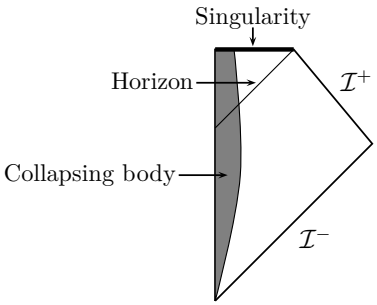


Figure 1: Penrose diagram for a collapsing spherical body. $\mathcal{I}^+$ and $\mathcal{I}^-$ denote past and future null infinity.

3 QFT on a fixed curved background: black holes radiate

The most dramatic change in our understanding of black hole physics came as a result of Hawking's famous analysis. What this analysis showed was that the formation of a black hole would modify the state of any quantum field in such a way that, at late times, there would be an outgoing flux of particles carrying energy towards infinity. Moreover, Hawking showed that the flux was characterized by the surface gravity κ of the resulting asymptotic stationary state of the black hole. This discovery transformed our perception of the formal analogy, originally pointed out in Bekenstein (1972), between the laws of black hole dynamics, and the standard laws of thermodynamics (see Wald (1994) for a discussion). In particular, it led to the view that the surface gravity is in fact a measure of the black hole's temperature $T = \frac{\kappa}{2\pi}$, and that the event horizon's area A is a measure of the black hole's entropy $S = A/4$.

Hawking's result is probably the most famous of the effects that arise from the natural extension of special relativistic quantum field theory to the realm of curved spacetimes. It imposes a dramatic modification on the classical view of black holes as

can escape to infinity.

absolutely black and eternal regions of spacetime. It is important to stress, though, that Hawking's calculation, being a result pertaining to quantum field theory on a *fixed* spacetime, does not encompass back-reaction effects. These are in fact notoriously difficult to deal with and a general framework for doing so is lacking. At any rate, some straightforward physical considerations, which have rather dramatic consequences, are often brought to bear in this context.

4 Back-reaction and first quantum gravity input: black holes evaporate

As can be expected, Hawking's result also suggests a dramatic modification in our expectation for the ultimate fate of a black hole. That is, while before Hawking's discovery, one would have expected that, once formed, a black hole would be eternal, the fact that the radiation is caring energy away, assuming overall energy conservation, leads one to expect that the mass of the black hole will start diminishing. The context in which this problem is standardly set is that of asymptotically flat spacetimes, for which we have a well defined notion of overall energy content given by the ADM mass[3] of the spacetime, a quantity which is known to be conserved.

As we noted, Hawking's calculation cannot deal with back-reaction. However, our confidence on energy conservation in the appropriate situations is so robust that it is difficult not to conclude that, as the radiation carries away energy, the black hole mass will have to diminishing. If this takes place, the surface gravity of the black hole—which is no longer really stationary, but can be expected to deviate from stationarity only to a very small degree—would change as well. As it turns out, the surface gravity is inversely proportional to the black hole's mass, so the black hole temperature can be expected to increase, leading to a ever more rapid rate of energy loss and a correspondingly faster decrease in mass.

The run away picture for the evaporation process suggests a complete disappearance of the black hole in a finite amount of time. Of course, we cannot really be sure about this picture because, in order to perform a solid analysis, we would need to deploy a, currently lacking, trustworthy theoretical formalism adept to the challenge. The

[3]The ADM mass is a quantity associated with the asymptotic behavior of the induced spatial metric of a Cauchy hypersurface. In asymptotically flat spacetimes, it is known to be independent of the hypersurface on which it is evaluated (see Arnowitt et al. (1962)).

problem is that, by the removal of energy from the black hole, one can expect to eventually reach a regime where quantum aspects of gravitation become essential to the description of the process. At such point, one might contemplate the possibility that, as a result of purely quantum gravitational aspects, the Hawking evaporation of the black hole will stop, leaving a small stable remnant. This, in turn, might open certain possibilities regarding the information issue. For the time being, though, we will ignore such an option.

Then, in order to simplify the discussion at this point, we will ignore the possibility of remnants and assume that there is nothing to stop the Hawking radiation. Then, if the black hole's mass decreases in accordance with energy conservation, one expects that the black hole to simply disappear and the spacetime region where it was located to turn flat (see Figure 2).

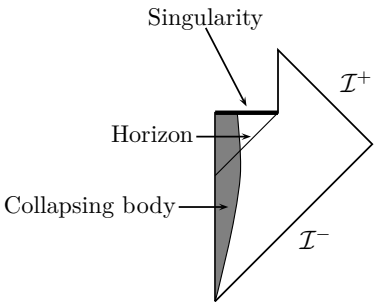


Figure 2: Penrose diagram for a collapsing spherical body, taking into account Hawking's radiation.

At this point, we seem to come face to face with an information loss problem: the original massive object that collapses, leading to the formation of a black hole, might have required an incredibly large amount of detail for its description. However, the final state that results from the evaporation is simply described in terms of the thermal Hawking flux, followed by an empty region of spacetime. More to the point, even if the initial matter that collapses to form a black hole was initially in a pure quantum state, after the complete evaporation of the black hole there would be a mixed one, corresponding to the thermal Hawking flux. These considerations seem to indicate that, even at the fundamental level, we have a fundamental loss of information. The final state, even if described in full detail, does not encode the information required to retrodict the details of the initial one. At the level of quantum theory, we would

be facing a non-unitary (and non-deterministic) relation between the initial and final states of the system, a situation that seems at odds with the unitary evolution provided by the Schrödinger equation.

There are, however, various caveats to the above conclusion. The first one is opened up by the possibility of the evaporation eventually stopping, leading to a stable remnant. The mass of said remnant can be estimated by considering the natural scales at which the effects of quantum gravity are expected to become important. This leads to an estimate of the order of Plank's mass ($\approx 10^{-5}$ gr). Then, if one wants the remnant to encode all the information present in the initial state, one is led to the conclusion that such a small object would have a number of possible internal states as large as that of the original matter that collapsed to form the black hole, which can, of course, have had a mass as large as one can imagine. It is hard, then, to envisage what kind of object, with such rather unusual thermodynamical behavior, would this remnant have to be. For this reason, this possibility is usually not considered viable (although we acknowledge that these considerations might be overturned; for a discussion of these issues see Banks (1994)). At any rate, we will not consider this possibility any further.

We should also mention another proposal which uses the idea that, while curing singularities, quantum gravity might open paths to other universes, which could be home to the missing information. Such information would be encoded either in a new universe or in correlations between it and ours. Besides the dramatic ontological burden, such proposal leaves open the possibility of these alternative universes emerging even in ordinary processes (which could, e.g., involve virtual black holes), leading to information loss in such standard scenarios. Alternatively, the information could be preserved, but impossible to retrieve in principle. We will also not consider this possibility any further.

A much more important caveat is the following: we have very solid results indicating that, associated with the formation of a black hole, there is always a singularity of spacetime appearing withing it. The strongest results in this regard are a series of theorems proved by Hawking (see Hawking and Ellis (1973)) showing that, under quite general conditions, and assuming reasonable properties for the energy and momentum of the collapsing matter, the formation of singularities is an inevitable result of Einstein's equations. The issue is that, at the classical level, these singularities represent a breakdown of the theory and, in fact, a failure of the spacetime description. The singularities are, therefore, to be thought of as representing boundaries of spacetime, rather that points within it. Once a spacetime has additional boundaries, it is clear

that the issue of information has to be confronted on a different light. Of course, if one considers the description of the system at an initial Cauchy hypersurface and wants a final hypersurface to encode the same information, one has to make sure that the final one is also Cauchy.

The formation of singularities then implies that, if we want to have spacetime regions where the system's state could be thought of as encoding all the information, then we must surround the singularities by suitable boundaries. In other words, if the singularities force us to include further boundaries of spacetime, then the comparison of initial and final information has to be done between the initial Cauchy hypersurface and the late-time *collection* of surfaces that, together, act as a Cauchy hypersurface. That collection could naturally include asymptotically null future, but also the hypersurfaces surrounding the singularities. The same kind of calculation as the one done by Hawking would then show that all the information present on the initial hypersurface would also be encoded in the state associated with this late-time Cauchy hypersurface. That is, if we include the boundary of spacetime that arises in association with the singularity, then there is no issue regarding the fate of information. We conclude that, under these circumstances, still there is no information loss.

5 Second quantum gravity input: black holes do not involve singularities

As we noted above, singularities represent a breakdown of the spacetime description as provided by general relativity and thus indicate the need to go beyond such theory. The expectation among theorists is that quantum gravity is going to be the theory that cures these failures of classical general relativity, replacing the singularities by a description in the language appropriate to quantum gravity. This is, in fact, what occurs with various other theories that are known to be just effective descriptions of a physical system's behavior in a limited context, but that have to be replaced with a more fundamental description once the system leaves that regime. Think for instance of the description of a fluid by, say, the Navier-Stokes equations. We know that this description works very well in a large variety of circumstances, but that a breakdown of such description occurs, for instance, when there are shock waves or when other types of singularities are formed. However, under such circumstances, the underlying kinetic theory, including the complex inter-molecular forces, is expected to remain valid. The

point is that, just as in those cases, one expects the emergence of singularities in general relativity to indicate the end of the regime where the classical description of spacetime is valid and, therefore, where a quantum gravity description would have to take over (see Figure 3 and Ashtekar and Bojowald (2005) for details).

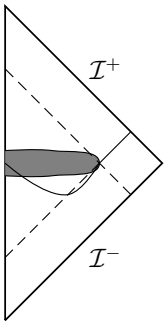


Figure 3: "Quantum spacetime diagram" for a black hole.

Of course, if quantum gravity does in fact cure the singularities, and removes the need to consider, in association with the corresponding regions, a boundary of spacetime, the issue of the fate of information in the Hawking evaporation of black holes resurfaces with dramatic force. So, do we finally have a genuine paradox in our hands. Not quite yet; a few elements are still missing. In order for a paradox to arise, we need to couple a genuine loss of information with a fundamental theory which does not allow for information to be lost.

6 A paradox?

When is it, then, that the Hawking radiation by a black hole leads to an actual paradox? We are finally in a position to enumerate the various assumptions required in order to construct a genuine conflict:

1. As a result of Hawking's radiation carrying energy away from the black hole, the mass of the black hole decreases and it either evaporates completely or leaves a small remnant.

2. In the case where the black hole leaves a small remnant, the number of its internal degrees of freedom is bounded by its mass in such a way that these cannot possibly encode the information contained in an arbitrarily massive initial state.

3. Information is not transfered to a parallel universe.

4. As a result of quantum gravity effects, the internal singularities within black holes are cured and replaced by something that eliminates the need to consider internal boundaries of spacetime.

5. The outgoing radiation does not encode the initial information.

6. Quantum evolution is always unitary.

We have already discussed the arguments in support of assumptions 1, 2, 3 and 4 and saw that, although by no means conclusive, they are reasonable. But what about 5 and 6? Well, in order to avoid a paradox, and assuming the first four assumptions to be true, at least one of them has to be negated. In order to explore the motivations and consequences of doing so, we must think clearly about how to interpret Hawking's calculation in a context in which 1, 2, 3 and 4 are the case.

As we remarked above, Hawking's calculation is performed in the setting of a quantum field theory over a fixed curved background. What one finds there is that an initial pure state of the field evolves into a final one which, when tracing over the inside region, reduces to a mixed thermal state. The key question at this point, then, is how to interpret such a final mixed state in a setting in which i) the black hole is no longer there, so there is no interior region to trace over, and ii) in which there is no singularity (or parallel universe) for the information to "escape into." As far as we can see, there are two alternatives: either one assumes that the mixed state arises only as a result of tracing over the interior region and maintains that the outgoing radiation somehow encodes the initial information—which amounts to negating 5; or one takes Hawking's result seriously and maintains that, even in this scenario, information is lost—which amounts to negating 6. Below we explore each option in detail.

6.1 The outgoing radiation encodes information

In the last couple of decades, the community's position on the information loss subject has been strongly influenced by developments in String theory. Such framework has permitted exploration of questions, regarding black holes, using settings where event horizons and singularities play no relevant roles. This is possible due to the AdS/CFT correspondence (see e.g., Strominger (2001)), which allows the mapping of complicated spacetime geometries in the "bulk" of asymptotically Anti-de Sitter spacetimes,

including ones involving black holes, onto corresponding states of an ordinary quantum field theory living on the Anti-de Sitter boundary (which is, in fact, a flat spacetime). These considerations have led people to conclude that, as a breakdown of unitarity is not expected to take place in the context of a quantum field theory in flat spacetimes, there should be no room for a breakdown of unitarity in the corresponding situation involving black holes either.[4]

The proposal, then, is that unitarity is never broken and that information is never lost. As a result, Hawking's calculation has to be somehow attuned to assure consistency. In particular, the proposal is that the outgoing radiation must encoded all of the initial information. There is, however, a high price to pay in order to achieve this. As has been shown in Almheiri et al. (2013), in order for the outgoing radiation to encode the necessary information, each emitted particle must get entangled with all the radiation emitted before it. However, due to the so-called, "monogamy of entanglement," doing so entails the release of an enormous amount of energy, turning the event horizon into a *firewall* that burns anything falling through it. The upshot then, is a divergence of the energy-momentum tensor of the field over the event horizon and a radical breakdown of the equivalence principle over such a region.

6.2 Unitarity is broken

The discovery of the Hawking radiation was initially taken as a clear indicative of information loss at the fundamental level. In fact, Hawking (1976) even introduced a notation for this general type of evolution which was supposed to account for the transformation from (possibly pure) initial states ρ_i into final mixed ones ρ_f. Hawking denoted the general linear, non-unitary, operator characterizing such transformation by the sign \$, i.e., $\rho_f = \$\rho_i$. Likewise, Penrose pointed out that, in order to have a consistent picture of phase space for situations involving black holes in thermal equilibrium with an environment, one has to assume that ordinary quantum systems undergo something akin to a self-measurement, by which he meant quantum state reduction that was not the result of measurement by external observers or measuring devices (see Penrose (1981)). Penrose (1999) further argued that quantum state reduction is probably linked to aspects of quantum gravity.

The early assessments of these ideas in Banks et al. (1984) indicated that they

[4]Note however that the argument can be easily reversed to show exactly the opposite. Since Hawking's result shows that unitarity breaks when black holes are present, one must conclude that quantum evolution *cannot* be unitary even in a quantum field theory on flat spacetimes.

where likely to lead to a very serious conflict with energy and momentum conservation or to generate unacceptable non-local features in ordinary physical situations. However, further analysis in Unruh and Wald (1995) showed that these assessments where not that solid and that there where various possibilities to evade the apparently damning conclusions.

In (omitted references) we have explored the viability of breaking unitarity both qualitatively and quantitatively. In particular, we have successfully adapted objective collapse models, developed in connection with foundational issues within quantum theory, in order to explicitly describe the transition from the initial pure state into a mixed one. Our view on the subject is based on the conviction that, contrary to the prevailing opinion in the community working on the gravity/quantum interface, there are good reasons to think that quantum theory requires modifications to deal with its basic conceptual difficulties. Below we discuss these issues and explore their consequences for the information loss paradox.

7 Information loss and the measurement problem

Most discussions of black holes and information loss do not implicate foundational issues of quantum theory. Of course, ignoring such issues, particularly with pragmatic interests in mind, is often acceptable. However, when deep conceptual questions are involved, such as in the present case, the pragmatic attitude might not be the right way to go.

The standard interpretation of quantum mechanics involves a profoundly *instrumentalist* character, with notions such as *observer* or *measurement* playing a crucial role. Such an instrumentalist trait becomes a problem as soon as one intends to regard the theory as a fundamental one, useful not only to make predictions in suitable experimental settings, but also to be applied to the measurement apparatuses, to the observers involved, or to non-standard contexts such as black holes or the universe as a whole. The resulting problem, often referred to as the *measurement problem*, has been discussed at length in numerous places and many different concrete formulations of it have been given. A particularly useful way to state it, given in Maudlin (1995), is as a list of three statements that cannot be all true at the same time:

A. The physical description given by the quantum state is complete.

B. Quantum evolution is always unitary.

C. Measurements always yield definite results.

Maudlin's formulation of the measurement problem is noteworthy because of its generality and its preciseness. Moreover, it is extremely useful in order to motivate and classify strategies to solve the problem. For example, by negating A, one arrives at so-called hidden variable theories, such as Bohmian mechanics; by removing B, one gets so-called objective collapse theories, such as GRW; and by discarding C, Everettian interpretations emerge. Of these three options, the last one is, by far, the most contentious. Among its most urgent matters, we can mention the problem of the preferred basis, the one of making sense of probabilities in the theory and the general and basic issue of establishing a clear and precise link between the abstract mathematical objects of the theory and concrete empirical predictions. Of course, brave attempts to deal with these and other issues within Everettian frameworks abound. However, be believe that, at least for the time being, they are far from being successful.

Returning to the measurement problem and its relation to the information loss issue, we note that assumptions 6 and B are in fact identical. Therefore, the strategy one decides to adopt in order to avoid complications regarding the information loss issue (e.g., negating 5 or 6 above) has implications with respect to what one must say regarding the measurement problem (e.g., negating A, B or C). In particular, if regarding the information loss, one decides to maintain the validity of 6 (and thus to hold that the outgoing radiation encodes all of the initial information), then one necessarily has to either negate A or C (i.e., either to entertain a hidden variables theory or an Everettian scenario). In other words, insisting on a purely unitary evolution, not only demands a violation of the equivalence principle and a divergence of the energy-momentum tensor, but also a commitment either with many worlds or with an acknowledgment that standard quantum mechanics is incomplete. On the other hand, if regarding the information loss problem, one decides to abandon unitarity, the same move automatically not only avoids a breakdown of the equivalence principle, but also guarantees success with respect to the measurement problem. The upper hand of the second option seems evident to us.

8 Conclusions

Since the publication of Hawking's analysis, more than forty years ago, the issue of black hole information loss has been a central topic in theoretical physics. The AdS/CFT

correspondence, proposed almost twenty years latter, came to further propel an already notorious debate. Yet, even after all these years, the discussion is often engulfed by confusion and misunderstanding among participants. The objective of this work is to develop a clear analysis of some of the key conceptual issues involved. Our hope is that, by doing so, significant progress on this important topic could soon be achieved.

We have presented the basic theoretical setting of the black hole information issue, paying special attention to elements, arising from not yet well-established physics, that presently have to be regarded merely as reasonable assumptions. Moreover, we have argued that the information loss issue is closely related to the measurement problem, and claimed that it is precisely within the context of certain proposals put forward to deal with the latter that the former finds one of its most conservative resolutions.

References

Almheiri, A., Marolf, D., Polchinski, J., and Sully, J. (2013). Black holes: complementarity or firewalls? *JHEP*, 62.

Arnowitt, R., Deser, S., and Misner, C. (1962). The dynamics of general relativity. In Witten, L., editor, *Gravitation: an introduction to current research*. Wiley.

Ashtekar, A. and Bojowald, M. (2005). Black hole evaporation: a paradigm. *Class. Quant. Grav.*, 22(3349).

Banks, T. (1994). Lectures on black hole information loss. *Nucl. Phys. Proc.*, 41.

Banks, T., Susskind, L., and Preskin, M. E. (1984). Difficulties for the evolution of pure states unto mixed states. *Nucl. Phys. B*, 224(125).

Bekenstein, J. D. (1972). Black holes and the second law. *Lett. Nuovo Cim.*, 4(737).

Hawking, S. W. (1976). Breakdown of predictability in gravitational collapse. *Phys. Rev. D*, 14(2460).

Hawking, S. W. and Ellis, G. F. R. (1973). *The large scale structure of spacetime*. Cambridge University Press.

Maudlin, T. (1995). Three measurement problems. *Topoi*, 14.

Penrose, R. (1981). Time asymmetry and quantum gravity. In Isham, C. J., Penrose, R., and Sciama, D. W., editors, *Quantum Gravity II*. Clarendon Press.

Penrose, R. (1999). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press.

Strominger, A. (2001). The AdS/CFT correspondence. *JHEP*, 0110(034).

Unruh, W. G. and Wald, R. M. (1995). On evolution laws taking pure states to mixed states in quantum field theory. *Phys. Rev. D*, 52:2176–2182.

Wald, R. M. (1994). *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics*. University of Chicago Press.

The Causal Homology Concept

Jun Otsuka*

Abstract

This presentation proposes a new account of homology, which defines homology as a correspondence of developmental or behavioral mechanisms due to common ancestry. The idea is formally presented as isomorphism of causal graphs over lineages. The formal treatment not only clears the metaphysical skepticism regarding the homology thinking, but also provides a theoretical underpinning to the concepts like constraints, evolvability, and novelty. The novel interpretation of homology suggests a general perspective that accommodates evolutionary developmental biology (Evo-Devo) and traditional population genetics as distinct but complementary approaches to understand evolution, facilitating further empirical and theoretical researches.

*Department of Philosophy, Kobe University, Rokko-dai 1-1 Nada, Kobe, Japan. Email: junotk@gmail.com

1 Introduction

The homology thinking, the idea that the same anatomical structure repeatedly appears in different species or parts of the same organism, has a long history in biology (Amundson, 2005). While the existence of such anatomical similarities among or within species is now explained by the descent from a common ancestor, the conceptual issues surrounding the notion have invited philosophical as well as methodological debates and skepticism. Owen famously defined homology as "the same organ in different animals under every variety of form and function," but this definition is perplexing rather than enlightening: what characterizes and warrants the sameness of "organs," if not their form or function? What, in other words, is the unit of homology?

There are three conceptual problems. The first and foremost problem is its *definition*: what exactly is homology? Evolutionary theory tells us that homology is identity due to a common origin, but an identity of *what*? Is it morphological characters, activities, clusters of properties, or genetic networks that are regarded to be same? And what is the criterion to judge whether or not two such things are actually the "same"? The second problem is *metaphysical*. As Ghiselin (1997) points out, the homology-as-identity partitions the whole tree of life into equivalence classes. But doesn't the supposition of such universal classes, reminiscent of Aristotelian essence, commit us to an anti-evolutionary thinking? And thirdly, there is a *pragmatic* question: why do we care about homology at all? Some neo-Darwinians such

as G. C. Williams see homologs as mere "residues," i.e. a relic of the past common ancestry not yet washed out by natural selection (Amundson, 2005, pp. 237-8). If that is the case homology by itself would have no explanatory role in evolutionary theory, and the quest for its definition, however well-defined and metaphysically sound, becomes a mere armchair exercise with no scientific value.

There is at least one usage of the concept free from these issues: homology of DNA sequences. Here the "sameness" is well-defined by matching bases that can be one of the four chemical kinds, G, C, T, A. Moreover, the scientific importance of orthologs and paralogs is undeniable in reconstructing the evolutionary history and predicting gene function, to name a few. Things become different for phenotype, in particular complex phenotypes like morphological or behavioral traits. First of all, there is no clear-cut definition of "phenotypic units" as that for nucleotides. Continuous traits such as height or weight usually lack objects breakpoints by which we classify them into discrete equivalence classes. In sum, there seem to be no non-arbitrary and non-controversial units for phenotype of which we can talk about the sameness, and thus homology.

Our first task, therefore, is to identify the units on which the phenotypic homology relationship can be defined. This presentation proposes that this purpose is best served by *causal graphs* which formally represent developmental or behavioral mechanisms. Homology is thus defined as graph isomorphism over lineages, or conservation of the underlying causal structure

over evolutionary history (Section 2). I will argue in Section 3 that the formal treatment of homology (i) solves the philosophical as well as empirical puzzles and criticisms regarding the homology concept; (ii) provides clear meanings to some key but elusive concepts such as constraints, evolvability, and novelty; (iii) and suggests a broad perspective that accommodates evolutionary developmental biology (Evo-Devo) and traditional population genetics as distinct but complementary research projects. Section 4 compares the present approach to other existing accounts of homology, and discusses its relative strengths, challenge, and philosophical implication. As will be stressed there, the primary objective of this presentation is to facilitate or open up new empirical as well as theoretical questions. The last section concludes with some of these research prospects that are prompted by the new homology concept.

2 Defining homology with graphs

The idea of characterizing homology in terms of causal structures is not new. Various biologists have suggested, albeit in different fashions, that the developmental or behavioral mechanisms underlying phenotype can or should serve as a unit of homology (e.g. Riedl, 1978; Wagner, 1989, 2014; Gilbert and Bolker, 2001; Müller, 2003). These proposal, however, are mostly based on independent examples or qualitative descriptions, and the lack of a unified treatment has blurred their philosophical as well as theoretical implications.

The aim of this section is to give a formal representation to the ideas of developmental sameness by using causal graphs, in view of exploring the conceptual nature of homology in the later sections.

A *causal graph* $\mathcal{G}$ is a pair $(\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a set of phenotypic or genetic variables of organisms and $\mathbf{E}$ is a set of edges representing causal relationships among these traits. Development is understood as a causal web connecting embryological, morphological, and behavioral traits, and the set of edges $\mathbf{E}$ characterizes these causal links. Note that such connections may remain invariant even under considerable modifications in phenotypic values or the functional form that determines the quantitative nature of each edge. The same set of $\mathbf{E}$ is consistent with a variety of phenotypic states and forms of causal production; it only defines the qualitative feature of the causal networks, i.e. which causes which.

Once modeled in this way, it becomes meaningful to compare causal structures of different organisms. A causal graph $\mathcal{G}_1 = (\mathbf{V}_1, \mathbf{E}_1)$ is *isomorphic* to another $\mathcal{G}_2 = (\mathbf{V}_2, \mathbf{E}_2)$ if they have the same structure, or more formally if there is a bijection $f : \mathbf{V}_1 \to \mathbf{V}_2$ such that if $(v, w) \in \mathbf{E}_1$ then $(f(v), f(w)) \in \mathbf{E}_2$. Likewise, isomorphism can be defined for subgraphs, which are just parts of the causal graphs restricted to a subset $\mathbf{V}' \subset \mathbf{V}$. We write $\mathcal{G}_1 \sim \mathcal{G}_2$ if two (sub)graphs are isomorphic. It is easy to see '$\sim$' is symmetric, reflexive, and transitive, and thus defines a equivalence class.

Each individual is assigned one causal graph that models a particular part of its developmental or behavioral mechanism. Let us denote the causal

structure of an organism a by $\mathcal{G}(a)$. Collectively, $\mathcal{G}(A)$ is a set of causal structures for a set of organisms A. We assume usual ancestor/descendant relationships over a set of organism Ω (which may include more than one species). If b is an ancestor of a, the *lineage* between b and a is a set of every individual between them. Given this setup homology is defined as follows.

> For two sets of organisms $A, B \subset \Omega$, let $\mathcal{G}'$ be a subgraph of all $g \in \mathcal{G}(A)$, and $\mathcal{G}''$ be a subgraph of all $g \in \mathcal{G}(B)$. Then $\mathcal{G}'$ and $\mathcal{G}''$ are homologous iff
>
> 1. $\mathcal{G}' \sim \mathcal{G}''$;
> 2. there is a set of common ancestors $C \subset \Omega$ of A and B[1]; and
> 3. for every d in all the lineages from C to A and C to B, $\mathcal{G}(d)$ has a subgraph $\mathcal{G}'''$ such that $\mathcal{G}''' \sim \mathcal{G}' \sim \mathcal{G}''$.

The definition explicates the idea that homology is the identity between causal structures due to common ancestry. Two (sets of) organisms share a homologous causal structure if, in addition to the graph isomorphism, every individual on the lineage connecting them shares the same causal graph, capturing the idea that the structure has been conserved through the evolutionary history.

The same treatment applies to serial homology, i.e. the homology relationship among parts of the same organism, such as teeth, limbs, or tree

[1]Note that C may be A or B themselves. Also note the condition 1 is redundant if a lineage includes the both ends. But here it is retained for clarity.

leaves. We can just set $A = B$, and compare different but isomorphic subparts $\mathcal{G}', \mathcal{G}''$ of the same overall structure $\mathcal{G}(A)$. Then the homology hypothesis is that there is an organism c in which the mechanism in question was duplicated, and the lineages from c to A have conserved the duplicated structures.

The above definition is illustrated with a case of special homology in figure 1, which depicts a particular region of the tree of life for (groups of) organisms A to G. Two mutations M_1, M_2 on the developmental mechanism occurred in the lineage leading to F, in which one causal edge $V_1 \to V_3$ was first removed and then restored. In this example, the causal structure $\mathcal{G}(D)$ of population D is homologous to $\mathcal{G}(E)$, for they are both inherited from the ancestral graph $\mathcal{G}(B)$ and $\mathcal{G}(A)$. In contrast, it is not homologous to $\mathcal{G}(F)$ even though they are graph-isomorphic. This is because the lineages connecting D and F do not conserve the causal structure in question: particularly it is not shared by C.

The example, though too simplistic to capture any real biological phenomena, makes explicit the idea that homology is a concordance of developmental mechanisms due to common ancestry. Note the criterion makes no reference to the resulting phenotype represented by particular values or distributions of variables. It does not require or forbid that, for example, two populations E and D show similar morphological distributions. Nor does it assume the graphs consist of the variables of the same nature. If the causal graphs in figure 1 represent a genetic network, kinds of genes/variables that

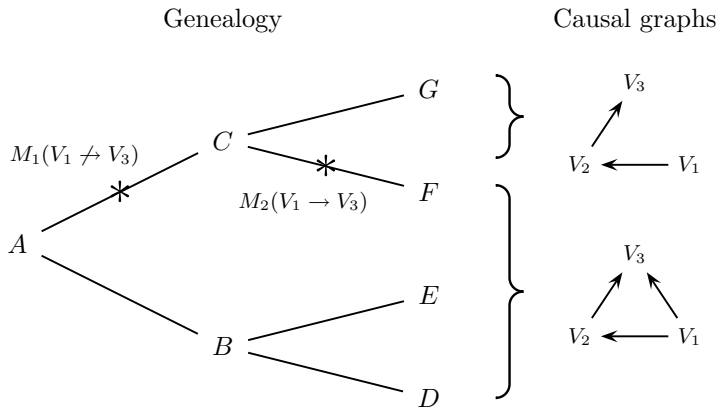


Figure 1: Illustration of graph homology. On the left is a genealogy tree for hypothetical populations A, B, C, D, E, F, G, while the graphs on the right describe causal structures of these populations over three characters, V_1, V_2, and V_3. Two asterisks ($*$) on the tree denote mutation events on the causal structure. See text for explanation.

constitute the network may vary across populations, as long as they serve the same causal roles within the overall structure. It is structural, rather than material, identity that defines homology. Theoretical as well as philosophical implications of this view will be explored in the following sections.

3 Conceptual advantages of the view

The above account is intended to provide a theoretical platform to formulate and evaluate hypotheses or explanations regarding homology. This section explicates the conceptual benefits of thinking homology in terms of causal graphs. Discussions on the empirical adequacy are differed to the next sec-

tion.

As discussed in the introduction, the major obstacle in defining homology is the absence of definite phenotypic units. Homology is an identity rather than similarity relationship (e.g. Ghiselin, 1997; Müller, 2003; Wagner, 2014), whereas no two or more phenotypic characters are identical in a strict sense — there are always subtle differences in, say, shape or size. The problem could be solved if we could find a natural and non-arbitrary way to factorize the phenotypic space into discrete regions so that two phenotypes within the same region are regarded "identical" despite their apparent differences. This is a difficult task, especially because we do not know the topological feature of the phenotypic space (Wagner and Stadler, 2003). To solve this issue the present analysis adopts a different strategy: instead of trying to impose a certain structure on the phenotypic space, it takes the generative mechanisms as basic units. Once these mechanisms are represented by causal graphs, which by nature are discrete mathematical entities, the desired identity relationship is given by graph isomorphism regardless of differences in the resulting morphology/phenotype. The graphical representation thus provides natural units prerequisite to define homology.

It is granted that a graph representation is not determined uniquely, because the same developmental mechanism can be modeled in various levels of abstraction, yielding causal graphs of different complexities. However, I take this to be a strength rather than weakness of my view, because homology too is often treated as description-dependent. Teleost fins and tetrapod limbs

are said to be homologous *as* paired vertebrate appendages, but *not as* fins or limbs. In contrast, our hands and pectoral fins of the whale are homologous not only as appendages but also as limbs. One tempting hypothesis is that such degrees of homology relationship correspond to isomorphisms of causal structures described at different granularities. In the above example, it is hypothesized that teleost fins and tetrapod limbs are represented by the same, but rather course-grained, causal graph, while tetrapod species share the causal structure to much finer details.

Fixing the level of abstraction determines not only the equivalent classes but also the degree of similarity between these classes. Two distinct causal graphs may be closer or further depending on the number of changes required to obtain one from the other. If $\mathcal{G}''$ is obtained by removing one edge from $\mathcal{G}'$ which in turn lacks one of the edges of $\mathcal{G}$, $\mathcal{G}''$ is one step further than $\mathcal{G}'$ from the original $\mathcal{G}$. Each such deletion or addition of causal connection is called *novelty*. Novelty in this framework is a modification of the causal graph, and as such creates a new equivalence class of causal graphs, namely homology. Evolutionary novelty also comes in different degrees. In general, a single modification in abstract graphs will correspond to multiple edge additions or deletions in detailed ones, and thus is weighted more. In this regard a change in the causal graph shared both by teleosts and tetrapods will count as a significant novelty and possibly a creation of a new "bauplan."

This brings us to one of the central contentions in today's evolutionary biology, namely the alleged inadequacy of the Modern Synthesis framework,

in particular population genetics, to incorporate macro-scale evolutionary phenomena uncovered by evolutionary developmental biology (e.g. Pigliucci and Müller, 2010). It has been claimed that homology (macro-scale conservatism) and novelty (a large phenotypic change) not only resist explanations by the Neo-Darwinian gradualism, but also constrain evolutionary trajectories as modeled in population genetics (e.g Amundson, 2005; Brigandt, 2007). The theoretical relationship between Evo-Devo and population genetics, however, remains elusive, which makes difficult to evaluate the call for the "new synthesis."

The present approach, by expressing homolgy and novelty in terms of graph equivalence and modification, suggests a perspective on this connection and a way to turn these claims into empirical hypotheses. Because causal models induce evolutionary changes as studied in population and quantitative genetics (Otsuka, 2015, 2016), the graphical representation allows one to analyze how developmental structures generate and constrain evolutionary dynamics. In particular, topological features of the graph such as modularity yield, via the so-called Markov condition, patterns of probabilistic independence on the phenotypic distribution and determine possible evolutionary trajectories or *evolvability*. The causal graph approach thus supports the view that a homolog constitutes a unit of morphological evolvability (Brigandt, 2007).

The graph structures that yield population dynamics are usually not study objects of population genetics. They rather serve as background frame-

works in which evolutionary models are build to study changes in genetic or phenotypic frequencies. These frameworks, however, must come from somewhere, and this evolutionary process is a primary interest of Evo-Devo. Studies on homology and novelty — graph stasis and change — amount to "higher order" evolutionary analyses that deal with changes in the theoretical framework used in population genetics to predict local population dynamics. The graphical conception of homology thus suggests a broad perspective that accommodates these different, and sometimes seen antagonistic, research fields as complementary approaches to understand evolution.

Finally, let us turn to the metaphysical problem. As seen above, homology is defined as an equivalence class over a set of causal graphs. But to what do such classes correspond, if not some ideal types or essences? Homology thinking has been criticized as anti-evolutionary due to its alleged commitment to essentialism. These critics thus re-interpret homology as a lineage that connects individual parts, rather than as a universal class to be instantiated by its members/homologs (e.g. Ghiselin, 1997). A detailed examination of this criticism must await another occasion, but here I just want to propose a different way to look at the issue. A metaphysical implication from the present study is that homology stands to concrete parts of organisms not as a universal to individuals, nor as a whole to parts, but rather as a model to phenomena to be modeled. A homology hypothesis is based on an observation that two or more individuals or parts thereof can be modeled by

the same causal graph.[2] Hence the proper relationship is not instantiation or mereology, but representation (Suppes, 2002). Once conceived in this way, the metaphysical ghost of essentialism vanishes away. Just like the same oscillator model characterizes various kinds of pendulum clocks, homology-as-model is a mathematical entity (directed graph) that may represent more than one actual individual, but that does not force us to commit to any form of essentialism.

The individual-universal distinction has also cast a shadow on the pragmatic issue regarding the epistemic role and significance of the concept of homology. It has been argued that the study of homology cannot be any more than a historiography since there is no such thing as a law for individuals (Ghiselin, 1997). A very different picture, however, emerges from the present thesis. A homology statement is a historical hypothesis regarding causal isomorphism — that two or more (sets of) organismal parts can be represented by the same causal model — and as such makes various predictions. For example, it supports extrapolations from model organisms, predicting that homologous organs will respond in the same or similar fashion to physiological, chemical, or genetic interventions. In addition, since isomorphic developmental structures will generate similar patterns of phenotypic variation (see above), their evolutionary changes are expected to follow similar trajectories. Establishing homologous relationships therefore is not a mere

[2]This, in turn, implies these individuals would respond in a more or less same fashion to hypothetical interventions (Woodward, 2003). Hence homology statements eventually boil down to counterfactual claims.

historical description, but has predictive implications both on physiological and evolutionary studies.

4 Comparisons and possible objections

This section compares the present proposal with some of the existing accounts of homology and also discusses possible objections. A number of philosophers and biologists have recently proposed to define homology as a *homeostatic property cluster*, a cluster of correlated properties maintained by "homeostatic mechanisms" (e.g. Boyd, 1991; Rieppel, 2005; Brigandt, 2009; Love, 2009). Since clustering and correlations are a matter of degree, homology according to this view is not an identity but a similarity relationship. It thus confronts with the boundary problem — to what extent properties must be clustered to form a homolog? The underlying "homeostatic mechanism" is supposed to clarify this boundary, but without a clear definition of what it is such an attempt only leads to a circularity. In particular, if it is defined as "those causal processes that determine the boundary and integrity of the kind (Brigandt, 2009, p.82)," the charge of circularity cannot be avoided.

This kind of problem will not arise if the generative mechanisms are defined explicitly in terms of causal graphs. While my approach proposes a formal framework to represent these mechanisms, it does not make any assumption or restriction on their structure: in particular it does not require the mechanism to be homeostatic, circumventing the criticism that a home-

ostatic mechanism by definition cannot evolve (Kluge, 2003). Moreover, the reference to "clusters" or even properties becomes superfluous, because the variational properties of phenotype are mere derivatives of the underlying causal graph. Of course, covarying traits suggest some ontogenetic connections, and thus may serve as a useful heuristics for finding homologs. They are, however, only "symptoms" — what *define* homology are not properties, clustered or homeostatic, but rather generative mechanisms.

The present approach has a closer affinity to the so-called *biological homology concept* that attempts to explain the phenomena of homology on the basis of a particular feature of the underlying causal structure, such as gene regulatory networks (e.g. Wagner, 1989, 2014). Indeed, one motivation of this presentation is to give a formal platform for these empirical hypotheses to elucidate their theoretical as well as philosophical implications. An important empirical challenge to the biological homology concept, and any other attempts to identify a homolog with a certain developmental structure, is the well-known fact that morphological similarity does not entail developmental sameness (Wagner and Misof, 1993). It has been reported that apparently homologous characters in related species may develop from different genes, cell populations, or pathways — the phenomena called *developmental system drift* (True and Haag, 2001). Although these phenomena present a challenge to my account as well, not all of them count as counter evidence. If, for example, "drift" concerns only genetic or cell materials, topological features of the causal network may remain invariant. Descriptive levels also matter.

Even if two causal structures differ at a fine-grained description, they may coincide at a more abstract level. Finally, my view does not require the entire developmental system to be conserved: if causal graphs share *some* part, they may still be homologous *in that aspect*. Indeed, it would be surprising if two apparent homologs turn out to share no developmental underpinnings at all. Some degree of flexibility may be expected, but so is inflexibility. Representing and comparing homologs in terms of the underlying causal graphs will serve as a heuristics to identify which part of the overall developmental system is responsible for generating similar morphological patterns.

From a philosophical perspective, a distinguishing feature of my account is its explicit reference to *models*. Homology has traditionally considered to be a relationship among concrete biological entities or properties thereof: it is organs or phenotypic features that are said to be homologous. In contrast, homology in my view is a relationship among abstract entities, i.e. causal graphs. How and why does such an abstract relationship reveal anything interesting about the concrete evolutionary history? That scientific theories and concepts should directly describe actual phenomena is a predominant view of science both in lay and scholarly circles. Under this conception logical positivists made it their primary task to define theoretical terms by the observable. In the same vein philosophers of biology have tried (not successfully in my view) to justify the concepts like homology or species by identifying necessary and sufficient conditions in terms of visible or directly verifiable features of organisms.

This apparently intuitive picture, however, has been criticized to be an overly simplistic view on the relationship between a scientific theory and reality (e.g. Suppes, 1967; Cartwright, 1983; Suppe, 1989). According to the critics the primary referents of scientific theories, concepts, and laws are not actual phenomena but idealized models. These models are not exact replicas of reality, but extract only certain features that are supposed to play essential roles in the scientific problem at hand. The present analysis is in line with this tradition. Causal graphs are highly idealized and thus possibly incomplete representations of complex causal interactions in living systems, but it is this idealization that affords explanatory power and general applicability. That is, on the condition that a model extracts the common causal structure of a population can it be used to predict the population's evolutionary trajectory or consequences of hypothetical interventions.

Most of these models, however, are still idiosyncratic to particular populations — e.g. population geneticists usually build, customize, or parameterize their model for each study object.[3] Homology thinking aims at even higher generality: its core idea is that some distinct species or organs allow for the same treatment/model in the analyses of their evolutionary fate or physiological performance. A homology statement is a historical hypothesis as to why such a unified explanation is possible at all. That is, it justifies the use of the same causal model based on evolutionary history, i.e. by the descent of the

[3]Models of adaptive evolution, however, may be extrapolated to the same or similar environmental conditions. In this regard, the analogical thinking and homological thinking represent two distinct ways to generalize evolutionary models.

causal graph from common ancestry. Hence homology is far from "residual," but has a significant explanatory value in biology — it allows an extrapolation of an evolutionary or physiological model to other contexts, and thus provides a basis for the highest-level generality in biological sciences.

5 Conclusion

The concept of homology presupposes phenotypic units on which identity relationships can be defined. The present analysis identified these units with causal graphs representing developmental or behavioral mechanisms and defined homology as graph isomorphism over lineages. The advantage of this formal concept is that it acknowledges the distinctive role of the study of homology while suggesting its connection to the traditional population genetics framework. That is, it not only provides definite meanings to such concepts like constraints, evolvability, and novelty, but also presents homology as a historical account or justification of the generalizability of evolutionary or physiological models. This is paralleled with the shift in the ontological nature of what can be said to be homologous: homology is a relationship between theoretical models, rather than concrete biological entities such as organs. Hence the proper relationship between homology to actual biological phenomena is not instantiation, but representation. Once conceived in this way the metaphysical problem of the alleged essentialism fades away.

The new account of homology prompts empirical, theoretical, and philo-

sophical researches on various topics, including the study of novelty and evolvability, the interplay between Evo-Devo and population genetics, implications of developmental flexibility, and the generalizability of biological models, to name a few. Another interesting philosophical question not mentioned above is the possibility of extending the current approach to another vexing concept in evolutionary biology, namely *species*. If homology is a partial matching of the causal structures between distinct species, it is tempting to define species by the whole causal structure — so that two organisms belong to the same species if their entire ontogeny and life history are represented by the same causal graph. This is a big question that requires an independent analysis, but will be briefly discussed in the presentation if time permitted.

References

Amundson, R. (2005). *The Changing Role of the Embryo in Evolutionary Thought: Roots of Evo-Devo*. Cambridge University Press, New York, NY.

Boyd, R. (1991). Realism, anti-foundationalism and the enthusiasm for natural kinds. *Philosophical Studies*, 61(1-2):127–148.

Brigandt, I. (2007). Typology now: homology and developmental constraints explain evolvability. *Biology & Philosophy*, 22(5):709–725.

Brigandt, I. (2009). Natural kinds in evolution and systematics: Metaphysical and epistemological considerations. *Acta Biotheoretica*, 57(1-2):77–97.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press, New York, NY.

Ghiselin, M. (1997). *Metaphysics and the Origin of Species*. State University of New York Press, New York.

Gilbert, S. F. and Bolker, J. A. (2001). Homologies of process and modular elements of embryonic construction. *Journal of Experimental Zoology*, 291(1):1–12.

Kluge, A. G. (2003). On the deduction of species relationships: A précis. *Cladistics*, 19(3):233–239.

Love, A. C. (2009). Typology reconfigured: From the metaphysics of essentialism to the epistemology of representation. *Acta Biotheoretica*, 57(1-2):51–75.

Müller, G. B. (2003). Homology: The Evolution of Morphological Organization. In Müller, G. B. and Newman, S. (eds.), *Origination of Organismal Form: Beyond the Gene in Developmental and Evolutionary Biology*, pp. 51–69. The MIT Press.

Otsuka, J. (2015). Using Causal Models to Integrate Proximate and Ultimate Causation. *Biology & Philosophy*, 30(1):19–37.

Otsuka, J. (2016). Causal Foundations of Evolutionary Genetics. *The British Journal for the Philosophy of Science*, 67(1): 247-269.

Pigliucci, M. and Müller, G. B. (2010). *Evolution: the extended synthesis*. MIT Press, Cambridge, MA.

Riedl, R. (1978). *Order in living organisms: a systems analysis of evolution*. Wiley, New York, NY.

Rieppel, O. (2005). Modules, kinds, and homology. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 304(1):18–27.

Suppe, F. (1989). *The semantic conception of theories and scientific realism*. University of Illinois Press.

Suppes, P. (1967). What is a scientific theory? In Morgenbesser, S. (ed.), *Philosophy of Science Today*, pp. 55–67. Basic Books, Inc., New York.

Suppes, P. (2002). *Representation and Invariance of Scientific Structures*. CSLI Publication, Stanford, CA.

True, J. R. and Haag, E. S. (2001). Developmental system drift and flexibility in evolutionary trajectories. *Evolution and Development*, 3(2):109–119.

Wagner, G. P. (1989). The biological homology concept. *Annu. Rev. Ecol. Evol. Syst*, 20:51–69.

Wagner, G. P. (2014). *Homology, Genes, and Evolutionary Innovation*. Princeton University Press, Princeton, NJ.

Wagner, G. P. and Misof, B. Y. (1993). How can a character be developmentally constrained despite variation in developmental pathways? *Journal of Evolutionary Biology*, 6(3):449–455.

Wagner, G. P. and Stadler, P. F. (2003). Quasi-independence, homology and the unity of type: a topological theory of characters. *Journal of theoretical biology*, 220(4):505–527.

Woodward, J. B. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York.

Serendipity: an argument for scientific freedom?

Baptiste Bedessem and Stéphanie Ruphy
Université Grenoble Alpes

PSA 2016, Atlanta

Abstract

The unpredictability of the development and results of a research program is often invoked in favor of a free, desinterested science that would be led mainly by scientific curiosity, in contrast with a use-inspired science led by definite practical expectations. This paper will challenge a crucial but underexamined assumption in this line of defense of scientific freedom, namely that a free science is the best system of science to generate unexpected results. We will propose conditions favoring the occurrence of unexpected facts in the course of a scientific investigation and then establish that use-inspired science actually scores better in this area.

1. Introduction

"I didn't start my research thinking that I will increase the storage capacity of hard drives. The final landscape is never visible from the starting point." This statement made by the physicist Albert Fert (2007), winner of the 2007 Noble Prize for his work on the giant magnetoresistance effect, expresses a very common belief, especially among scientists, about the unpredictable nature of the development and results of a research program. Such retrospective observations feed a type of "unpredictability argument" often invoked in favor of a pure, disinterested science led by scientific curiosity, in contrast with a use-inspired or applied science led by practical considerations. Polanyi gave a somewhat lyrical form of this kind of unpredictability argument in his classical essay "The Republic of Science" (1962). Science, says Polanyi (1962, 62), "can advance only by unpredictable steps, pursuing

problems of its own, and the practical benefits of these advances will be incidental and hence doubly unpredictable. … Any attempt at guiding research towards a purpose other than its own is an attempt to deflect it from the advancement of science… You can kill or mutilate the advance of science, but you cannot shape it." In Polanyi's view, claims about the unpredictable nature of scientific development go hand in hand with a plea for an *internal* definition of research priorities: a problem should be considered important in light of considerations internal to a field of scientific inquiry and not (at least not primarily) in light of external considerations, such as practical utility. The orientation of the inquiry by practical objectives is then deemed epistemically counter-productive and vain: one should not attempt to predict the unpredictable.

In response to this line of defense of free science, some authors emphasize the epistemic fecundity of use-inspired science (Stokes 1997, Wilholt 2006, Carrier 2004) showing that the presence of practical objectives does not run counter to the building of fundamental knowledge: more fundamental knowledge may be needed to achieve some particular practical ends. Industry research on the giant magnetoresistance effect in the 1990s is a telling example of research undertaken under considerable pressure to produce applicable results but which nevertheless produced, along the way, new fundamental knowledge (Wilholt 2006).

Our aim in this paper is to develop another line of defense of the epistemic fecundity of applied science, by challenging a crucial but often implicit assumption in the traditional defense of scientific freedom based on scientific unpredictability (such as Polanyi's or Fert's), namely the assumption that a free science is the best system of science to generate unexpected facts. But what are actually the conditions favoring the emergence of novelty in the course of a scientific investigation? This important issue has not received much epistemological

attention.[1] We will fill this gap by first distinguishing two kinds of unpredictability arguments often mixed when debating on scientific freedom, to wit, unpredictability as unforeseen practical applications and unpredictability as *serendipity* (cases, as we will explain in more details, where unexpected facts open up new lines of inquiry). Focusing on the latter, we will propose two conditions that favor the occurrence of unexpected facts in the course of a scientific investigation. In light of these two criteria we will then compare pure, disinterested science and applied science as regards their capacity to generate novelty.

2.Two types of unpredictability arguments

Appeals to the unpredictability of scientific results actually refer to various kinds of situations, which need to be clearly distinguished. First, the notion of unpredictability of scientific results can designate unforeseen practical applications of fundamental knowledge. Second, it can refer to a serendipitous dynamics of scientific progress: a line of research may sometimes lead to a totally unexpected, surprising result, which opens a new direction of inquiry. These two kinds of unpredictability give rise to *distinct* arguments in favor of scientific freedom, unfortunately often mixed in discussions about the relative merits of pure science and application oriented science.

2.1 Unpredictability as unforeseen practical applications

When unpredictability refers to unexpected applications, the argument is the following: freedom of research should be preserved since a free, disinterested science is needed to generate a reservoir of fundamental knowledge, which then can be used to develop

[1] Wilholt and Glimell (2011, 353) do touch upon this issue when discussing the link made by proponents of the autonomy of science between freedom of research and diversity of approaches favoring the epistemic productivity of science. But they just note that it is a strong assumption and do no further discuss its validity.

applications. This argument was typically developed by Vannevar Bush who appealed to the now classically called linear model of innovation:

> "Basic research leads to new knowledge. It provides scientific capital. It creates the fund from which the practical applications of knowledge must be drawn. New products and new processes do not appear full-grown" (1945, 20).

The development of the H-bomb in the frame of the Manhattan project is a paradigmatic case, also invoked by Bush: "basic discoveries of European scientists" (1945, 20) about the structure of the matter is what made possible the military application. Another frequently cited example of unpredictable application is the invention of the laser, a widely-used technological device nowadays, made possible by pure theoretical developments in quantum physics during the first half of the XXth century.

We will not in this paper discuss further this first version of the unpredictability argument. Let us just mention that its underlying linear model of innovation linking pure science and practical applications has already been challenged on several grounds by various authors (e.g. Brooks, 1994; Leydesdorff, 1997; Edgerton, 2004; Rosenberg, 1992). We rather want to focus on the second (and also widespread) type of unpredictability arguments, whose validity has been much less scrutinized.

2.2 Unpredictability as serendipity

This second type of argument appeals to unpredictability in the sense of *serendipity*: an unexpected observation or result opens up a new line of research leading to a fundamental discovery. A very well known historical episode illustrating such a serendipitous scientific dynamics is the invention of the first antibiotic by Flemings, after he had accidentally

observed the effect of a fungi (Penicilium) on bacteria colonies (Flemings, 1929). Also often cited is the discovery of radioactivity by Henri Becquerel (1896): when working with a crystal containing uranium, Becquerel noted that the crystal had fogged a photographic plate that he had inadvertently left next to the mineral. This observation led to the hypothesis that uranium emitted its own radiations. Another, perhaps less cited instance of serendipitous scientific dynamics is the discovery of the chemotherapeutic cisplatine molecule by scientists initially working on the effects of an electric field on bacteria growth (Rosenberg *et al*.,1967). They observed that cell division was inhibited because of the unexpected formation of a chemical compound with the Platinum atoms contained in the electrode. This chemical compound, which they named cisplatine, was then successfully tested as an anti-proliferative agent against tumoral cells.

When unpredictability refers to such serendipitous discoveries, freedom of research is defended on the grounds that scientists should be able to freely change the direction of their research or open up new lines of inquiry, in order to be able to follow up on unexpected results, thereby generating new knowledge (which in turn will possibly lead to new applications). But to properly work as an argument favoring free, disinterested research over applied research, this "serendipity argument" actually presupposes that the occurrence of surprising facts is more likely to happen in the first system of science than in the second. For increasing the production of new knowledge (and possibly new applications) does not only depend on being able to freely follow up on unexpected facts, it also (obviously) depends on whether occurrences of unexpected facts are favored, to start with. Two types of considerations are thus mixed in the serendipity argument: considerations on the occurrence of unexpected facts and considerations on the (institutional, material) possibility to follow up on them.

We will not for the moment discuss the second type of considerations and focus on the first, which has been largely neglected in the literature on scientific freedom, namely the conditions that favor the occurrence of surprising facts. Our central issue is thus the following: is a use-inspired science less likely to generate unexpected results than a free science mainly fuelled by curiosity? After having clarified the notion of *unexpected* result, we will propose two criteria that, we will argue, favor the occurrence of such results and in light of which free science and applied science can be compared.

3. Conditions of emergence of unexpected facts

By "unexpected facts" occurring in the course of an inquiry, we simply mean here results (observations, outcomes of an experiment, etc.) that cannot be accounted for within the theoretical or, more largely, the epistemic framework in which the empirical inquiry has been conceived and conducted. This kind of "exteriority" is what leads scientists to move away from the initial explanatory framework and open up new lines of inquiry in search of an alternative one that could accommodate the unexpected results.

3.1 Isolation and purification of phenomena

It is now a well-known feature of contemporary experimental sciences that many of their objects under study are "created" in the laboratory rather than existing "as such" in the real world. When drawing our attention to this epistemologically important feature, Hacking (e.g. 1983, chap. 13) specified that we should not read this notion of "creation" of phenomena as if *we* were *making* the phenomenon, suggesting instead that a phenomenon is "created" in the laboratory to the extent that it does not exist outside of certain kinds of apparatus. This is typically the case for a phenomenon like the Hall effect: it did not exist "until, with great ingenuity, [Hall] had discovered how to *isolate*, *purify* it, create it in the laboratory" (Hacking

1983, 226, *our italics*). In other words, Hall created in 1879 the material arrangement - a current passing through a conductor, at right angles to a magnetic field –, for the effect to occur and "if anywhere in nature there [were] this arrangement, *with no intervening causes,* then the Hall effect [would] occur" (1983, 226, *our italics)*. Isolation, purification, control of intervening causes (i.e. control of physical parameters) are noticeable features of an experimental protocol that have a straightforward consequence directly relevant for our philosophical interrogation on serendipity: they tend to limit the number of causal pathways which can influence the response of the object or phenomenon under study experimentally. Unknown causal pathways existing in the real world are thus inoperant (or less operant) in laboratory conditions, thereby limiting the occurrence of unexpected results. Hence our first criterion to evaluate whether a certain system of science favors surprising results: the more the phenomena under study in that system are isolated, purified in highly regimented experimental conditions, the less likely the occurrence of unexpected results is.

Moreover, isolation, purification of phenomena often go hand in hand with another noticeable feature of laboratory sciences, described by Hacking as follows: "as a laboratory science matures, it develops a body of types of theory and types of apparatus and types of analysis that are mutually adjusted to each other" (1992, 30). In particular, a given theoretical framework determines the type of questions that can be probed experimentally, guides the design of apparatus and defines the type of data produced. Consequently, "data uninterpretable by theories are not generated" (Hacking 1992, 55). This process of mutual constraints is well illustrated for instance by recent experimental inquiries in particle physics, such as the quest for the Higgs Boson. Its existence was postulated in the frame of the Standard Model of theoretical physics (Higgs, 1964) and complex experimental apparatus have been developed with the explicit goal of "discovering" it (LEP, 2003). The "discovery" occurred in 2012 (ATLAS, 2012) but the high degree of tailoring of the apparatus to the

theory postulating the particle can be considered as imposing some kind of a priori structure on the phenomenon, so that particles such as the Higgs boson are not so much "discovered" than "manufactured" (Falkenburg, 2007, 53). In any case, the "discovery" of the Higgs boson was hardly a surprise and illustrates Hacking's more general contention about experimental inquiries typical of contemporary laboratory sciences as opposed to real-world experiments: "[their] results are more often *expected* than *surprising*" (1992, 37, *our italics*).

3.2 Theoretical unifying ambition

Another relevant characteristic of an experimental inquiry is the degree of generality of its theoretical framework. Scientists working within a theoretical framework with a large unifying scope will be reluctant to "leave" it and search for an alternative one when facing an unexpected result, and for good epistemological reasons: there is (obviously) a high epistemic cost of abandoning a theoretical framework that provides explanations for a large set of phenomena. The right move is rather to try to accommodate the surprising result by adopting, if necessary, *ad hoc* hypothesis or tinkering with some ingredients of the existing theoretical framework, so that the result looses its "exteriority" and ends up being integrated. And because of this well-known "plasticity" and integrative power of well-established theoretical frameworks with a large unifying scope[2], when a (at first sight) surprising result occurs, it rarely leads to the opening up of a new line of inquiry in search of an alternative explanatory framework, but rather gets integrated within the existing one, thereby losing its unexpectedness.

There is another reason why a high degree of theoretical generality does not favor the occurrence of unexpected results, which is linked to our previous remarks on the process of

[2] Classical references on these ideas of plasticity or integrative power are of course Kuhn's description (1962) of scientists being busy working on resolving anomalies in normal science and Lakatos' concept of "protective belt" of a research program (1978).

mutual adjustment between theoretical ingredients, apparatus and data. By constraining the type of experimental procedures developed and the type of data generated, a theoretical framework with a large unifying scope tends to *homogenize* the experimental works conducted to probe the various phenomena that it accounts for. And since a diversity of experimental approaches increases the possible sources of emergence of surprising facts, we can conclude that by reducing this diversity, theoretical generality makes the occurrence of unexpected facts less likely to happen.

The case of the etiology of cancer provides interesting illustrations of these two unexpectedness-diminishing effects of theoretical generality. The classical theory of cancer, the Somatic Mutations Theory (SMT), has been challenged for fifteen years or so by a new theoretical approach, the Tissue Organization Field Theory (TOFT) (Sonnenschein and Soto, 2000). First developed in the 1970's, the SMT rapidly became the dominant research theoretical framework on carcinogenesis (Mukherjee, 2010). This hegemony led to a high degree of homogenization of the experimental inquiries: the experimental procedures were all dedicated to the very standardized search for genetic mutations, in the context of molecular biology. Moreover, many, if not all surprising observations were made compatible with SMT by using *ad hoc* hypothesis (Soto, 2011). For instance, it was observed that various types of cancer were exhibiting large-scale disorganization of the genome. This observation was unexpected to the extent that it could not fit with SMT's fundamental postulate of punctual mutations. To integrate it in the frame of SMT, the existence of an original genetic instability of the cancer cells was then postulated (Rajagopalan, 2003).

4 Use-inspired science, pure science, and unexpected facts

In light of the criteria that we proposed above, how does pure, disinterested science score compared to applied science when it comes to favoring the occurrence of unexpected facts? A

helpful starting point is provided by Martin Carrier's insightful characterization of applied science:

> "Three methodological features can be observed whose combined or marked appearance tends to be characteristic of applied science: local models rather than unified theories, contextualized causal relations rather than causal mechanisms, real-experiments rather than laboratory experiment conducted for answering theoretical questions" (2004, 4).

4.1 Local models

Let us start with the contrast between local models and unified theories. Whereas pure science often aims at providing comprehensive and unifying theoretical frameworks (think of the Standard Model in particle physics or the Big Bang model in cosmology), use-inspired research is characterized by the coexistence of numerous local models, each determining the development of specific experimental procedures. An extreme case of this locality are for instance the design-rules used in the industry, which are built as laws guiding action (Wilholt, 2006). They are experimentally confirmed rules providing relations among different relevant parameters to manufacture industrial products. These rules are extremely specific: they apply to a very few number of situations and each of them determines a singular experimental practice. The use of local models is also widespread in the biomedical sciences, a typically use-inspired field of research. We will again draw on oncology to illustrate our point. Consider for instance the case of the development of radiotherapy protocols in the first half of the XXth century. The aim was to intervene on cancer to cure it, without any general model describing the mechanism of carcinogenesis. This program promoted the development of a variety of exploratory approaches using X-rays against cancer (Pinell, 1992). As there were

no standardized protocols, many experimental procedures were tested, changing the density of X-rays received, the distance of emission, the frequency of the radiotherapy sessions. In order to improve the efficiency of the therapeutic methods, scientists tried to build various local models describing the action of X-rays on cancer, corresponding to the variety of experimental procedures implemented. Grubbe (1949) formulated a model based on the inflammatory reaction to explain the effects of radiotherapy on cancer: the inflammation of the surrounding tissue beyond the effects of X-rays is responsible for the decrease of tumoral mass. This model is applicable to his specific use of X-rays: he applied very high doses, necessary to generate an inflammatory response. In parallel, Tribondeau and Bergonié, using more moderate doses, developed a model based on the proliferation of the cells in tumoral context, which led to the "Bergonié law": X-rays have a higher impact on proliferating cells (Tribondeau, 1959).

What lessons can be drawn from this first contrast between local models and unified theories? The answer is rather straightforward, given the link spelled out in the previous section between the level of generality of theoretical models and the occurrence of unexpected facts (our second criterion): by promoting the use of a diversity of local models and heterogeneous experimental protocols, applied science favors the occurrence of unexpected facts, whereas the penchant of pure science for comprehensive unifying theoretical frameworks, hence homogenized experimental protocols, does not.

4.2 Causal incompletness

Let us compare now pure science and applied science in light of our first criterion based on the degree of isolation and purification of the phenomena under study. A directly relevant feature of applied science is the use of what Carrier calls “contextualized causal relations” rather than full causal chains. Use-inspired science typically aims at directly intervening on a

process or phenomenon often disposing only of a partial knowledge of the causal chains involved and without being able to isolate it from various causal influences exerted by the rest of the physical world. A direct consequence of this feature of applied science is the low degree of control of its experimental protocols. By contrast, since pure science aims primarily at answering fundamental theoretical questions, it designs highly regimented experimental procedures that isolate and purify phenomena in order to be able to get empirical answers about the specific fundamental processes questioned in the theoretical investigation[3]. Moreover, building highly regimented experimental procedures requires knowledge of full causal chains in order to be able to better control the response of the system under study. The outcome of the application of our criterion is then again straightforward: compared with pure science, applied science favors the occurrence of unexpected facts to the extent that its experimental procedures are less controlled and based only on partial knowledge of the causal influences exerted on the phenomenon under study.

The etiology of cancer provides again interesting illustrations of our claim. Indeed, many current cancer therapies built in the frame of use-inspired research are based on contextualized causal relations. Typically, if a cellular agent is found to be massively expressed in cancer cells, drugs are designed to inhibit it, even if the whole causal chain determining its action is not known. For instance, a large amount of proteins promoting angiogenesis (the growth of blood vessels), notably VEGF (Vascular Endothelial Growth Factor), was found in tumoral cells, leading to the design of anti-VEGF molecules (Sitohy,

[3] Carrier sums up this contrast as follows: "Empirical tests often proceed better by focusing on the pure cases, the idealized ones, because such cases typically yield a more direct access to the processes considered fundamental by the theory at hand. But applied science is denied the privilege of epistemic research to select its problems according to their tractability (...). Practical challenges typically involve a more intricate intertwinement of factors and are thus harder to put under control". (2004a, 4) In the life sciences, this focus on "pure cases" means using "model organisms" or a limited number of well spread cell lines (e.g. the HeLa cells or the Saccharomyces Cerevisiae yeast) to elucidate fundamental biological mechanisms. And the use of such standardized objects tends to homogenize the experimental protocols.

2012). These molecules are used without considering the complete causal chain in which the VEGF is embedded. Only their known action on angiogenesis is considered. The clinical tests have led to unexpected observations: the use of an anti-VEGF molecule (Avastin) can stimulate tumor growth (Lieu *et al*., 2013)[4]. This example shows that the use of contextualized causal relations promotes the appearance of surprising facts by allowing unknown mechanisms to intervene in the experimental procedure.

5. Concluding discussion

Our previous analysis has established that several features of pure, disinterested science make it less hospitable than use-inspired science to the occurrence of unexpected facts. For all that, it does not follow that proponents of freedom of science cannot appeal anymore to unpredictability in the sense of serendipity to make their case. For the issue of which conditions favor the occurrence of unexpected facts is only half of the story. The other half is the possibility to actually follow up on these occurrences and open new lines of inquiry. And this other half raises different issues. What are the institutional, social structures of science that make it easier for scientists to re-orient their research when needed? To what extent an initial orientation of a scientific investigation by "external" practical needs is less compatible with the opening of new lines of inquiry than an initial orientation by epistemic considerations internal to the dynamics of a scientific field? When appealing to the serendipity argument,

[4] Interestingly, this observation led to new use-inspired research programs, aiming at identifing the molecular causal pathways giving rise to this tumoral resistance phenomenon. It has notably strongly oriented the research toward the precise understanding of the VEGF pathways (Moens, 2014). For instance, the study of the mechanisms of expression in cancer cells of various kinds of VEGF agents is becoming an important program of research (Li, 2014) and these works allow to build new fundamental knowledge about the action of the VEGF proteins.

proponents of free, disinterested science not only presuppose that it is the best system of science to generate unexpected facts to start with – a contention that we have challenged in this paper – but also that it actually gives more freedom to scientists to follow up on unexpected results. In other words, the issue of scientists' given possibility to change the direction of their research when needed is somewhat mixed, confused with the normative issue of what the aims of science should be (in short, increase knowledge following considerations internal to science *vs*. answer external practical needs). But it seems to us that the two issues should be kept separate. After all, one can very well conceive a system of science whose aims are primarily to answer society needs but which nevertheless leaves scientists free to choose the lines of inquiry that seem *to them* the most promising ways of fulfilling these needs (which includes changing research directions if needed). Otherwise put, one can very well conceive a use-inspired science which is not a *programmed* science in which scientists are asked to plan every step of their inquiry in order to achieve a given aim. And note that a pure, disinterested science may be as much programmed as a use-inspired science: the fact that scientists are left free to choose the aims of their research does not protect them from having to plan every step to reach these aims. In any case, our purport in this paper was not to attack pure, disinterested science. There are, no doubt, many good reasons to defend it, but the widespread, traditional one grounded on the unpredictability of scientific inquiry is certainly not the most epistemologically cogent and solid one.

REFERENCES

ATLAS Collaboration. 2012. "Obervation of a new particle in the the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". *Phy. Lett. B* 716(1) :1-29

Becquerel, Henri. 1896. "Sur les radiations émises par phosphorescence". *Comptes-rendus de l'Académie des sciences*.

Brooks, Harvey. 1994. "The relationship between science and technology". *Research Policy* 23(5):477-86

Bush, Vannevar. 1945. *Science, The Endless Frontier. A Report to the President by Vannevar Bush, Director of the Office of Scientific Research and Development*. Washington D. C.: National Science Foundation.

Carrier, Martin. 2004. "Knowledge and Control: On the Bearing of Epistemic Values in Applied Science". In *Values and Objectivity in Science*, ed. P. Machamer and G. Wolters, 275-293. Pittsburgh, PA: University of Pittsburgh Press.

———. 2004a. "Knowledge gain and practical use: Models in pure and applied research". In *Laws and Models in Science*, ed. D.Gillies, 1:17. London: King's College Publications

Edgerton, David. 2004. "The Linear Model Did not Exist. Reflections on the History and Historiography of Science and Research in Industry in Twentieth Century". In *Science-Industry Nexus: History, Policy Implications,* ed. Karl Grandin and Nina Wormbs, 31-57. New-York: Watson.

Falkenburg, Brigitte. 2007. *Particles Metaphysics. A critical Account of Subatomic Reality*. Springer.

Fert, Albert. 2007. Interview published in *Le Monde,* October, 25, 2007.

Flemings, Alexander. 1929. "On the antibacterial action of cultures of a penicillium with special reference to their use in the isolation of b. influenza". *J. Exp.Path*. 10:226-36.

Grubbe, Emil. 1949. *X-Ray Treatment: Its Origins, Birth, and Early History*. St.Paul and Minneapolis, MN: Bruce Publishing Company.

Hacking, Ian. 1983. *Representing and Intervening*. Cambridge University Press.

———.1992. "The Self-Vindication of the Laboratory Sciences". In *Science as*

practice and culture, ed. A. Pickering, 29-64. The University of Chicago Press.

Higgs, Petter W. 1964. "Broken Symmetries and The Masses of the Gauges Bosons". *Phy.Rev. Lett* 13:508.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. The University of Chicago Press.

Lakatos, Imre. 1978. *The methodology of scientific research programs*. Cambridge: Cambridge University Press.

LEP Collaboration. 2003. "Search for the Standard Model Higgs Boson at LEP". *Physics Letters B* 565:61-75

Leydesdorff, Loet and Etzkowitz Henry. 1997. *Universities and the Global Knowledge Economy: A Triple Helix of University-Industry-Government Relation*. London, Cassel Academic.

Li, Dong et *al*. 2014. "Tumor resistance to anti-VEGF therapy through up-regulation of VEGF-C expression". *Cancer Lett* 346:45-52.

Lieu, Christopher H. *et al*. (2013). The association of alternate vegf ligands with resistance to anti-vegf therapy in metastatic colorectal cancer. *PLoS One* 8(10):e77117.

Moens, Stijn et *al*. 2014. "The multifaceted activity of VEGF in angiogenesis - Implications for therapy responses". *Cytokine Growth Factor Rev* 25:473-82.

Mukherjee, Siddhartha. 2010. *The Emperor of All Maladies. A Biography of Cancer*. * Scribner.

Pinell, Patrice. 1992. *Naissance d'un fleau. Histoire de la lutte contre le cancer en France (1890-1940)*. Métaillé.

Polanyi, Michael. 1962. "The Republic of Science: Its Political and Economic Theory". *Minerva* 1: 54-73.

Rajagopalan, Harith, Nowak Martin A, Vogelstein Bert and Langauer Christoph. 2003. "The

significance of unstable chromosomes in colorectal cancer". *Nat Rev Cancer* 3(9):695-701.

Rosenberg, Nathan. 1992. "Science and Technology in the Twentieth Century". In *Technology and Enterprise in Historical Perspective*. Oxford, Clarendon Press.

Rosenberg, Barnett *et al*. 1967. "The inhibition of growth or cell division in escherichia coli by different ionic species of platinum(iv) complexes". *J. Biol.Chem* 242(6):1347-5.

Sitohy, Basel. 2012. Anti-vegf/vegfr therapy for cancer: reassessing the strategies. *Cancer Res* 8:1909-14.

Sonnenschein, Carlos and Soto A- M. 2000. "Somatic mutation theory of carcinogenesis: why it should be dropped and replaced". *Mol. Carcinog*. 29(4):205-211.

Soto, Ana M. and Sonnenschein C. 2011. "The tissue organization field theory of cancer: a testable replacement for the somatic mutation theory". *Bioessays* 33(5):332-340.

Stokes, Donald E. 1997. *Pasteur's Quadrant. Basic Science and Technological Innovation*. The Brookings Institution.

Tribondeau, Jean B. 1959. "Interpretation of some results of radiotherapy and an attempt at determining a logical technique of treatment". *Radiation Research*, 11(4):587-588.

Wilholt, Torsten. 2006. "Design rules: Industrial research and epistemic merit". *Philosophy of Science* 73(1):66-89.

Wilholt, Torsten and Glimell H. 2011. "Conditions of Science: The Three-Way Tension of Freedom, Accountability and Utility". In *Science in the Context of Application*, eds. M.Carrier and A.Nordmann, 351-70. Boston Studies in the Philosophy of Science.

PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association -460-

(Accepted for publication in *Philosophy of Science,*
subject to revision after presentation at 2016 PSA meeting)

Using Democratic Values in Science: an Objection and (Partial) Response[1]

S. Andrew Schroeder (aschroeder@cmc.edu),
Claremont McKenna College

draft of June 2016

Abstract

Many philosophers of science have argued that social and ethical values have a significant role to play in core parts of the scientific process. A question that naturally arises is: when such value choices need to be made, *which* or *whose* values should be used? A common answer to this question turns to political values — i.e. the values of the public or its representatives. In this paper, I argue that this imposes a morally significant burden on certain scientists, effectively requiring them to advocate for policy positions they strongly disagree with. I conclude by discussing under what conditions this burden might be justified.

1. Values in Science and the Political View

By now, most philosophers of science probably agree that there is an important place for so-called contextual (i.e. personal, ethical, political) values in core parts of the scientific process, especially in areas where science is connected to policy-making. Values may appropriately play a role in evaluating evidence (Douglas 2009), choosing scientific models (Elliott 2011), structuring quantitative measures (Reiss 2013, ch. 8; Stiglitz, Sen, and Fitoussi 2010; Hausman

[1] For comments on earlier drafts of this paper, I thank Alex Rajczi and the students in a seminar on science and values at Claremont McKenna College. For discussions on related topics, I thank Gil Hersch, Daniel Steel, and Branwen Williams. This work was supported in part by a research grant from the Claremont McKenna College Center for Innovation and Entrepreneurship.

2015), and/or in preparing information for presentation to non-experts (Elliott 2006; Hardwig 1994; Resnik 2001; Schroeder 2016). The natural follow-up question has received less sustained attention: when scientists should make use of values, *which* (or *whose*) values should they use?[2]

In some cases, philosophers of science criticize a value choice on substantive ethical grounds (e.g. Shrader-Frechette 2008; Hoffmann and Stempsey 2008). This suggests that the values to be used are the objectively correct ones. A second common view gives scientists latitude to choose whatever (reasonable) values they prefer or think best, usually supplemented by a requirement of transparency. This is suggested by many existing codes of scientific ethics, which impose few constraints on scientists in making such choices.[3] Finally, a third view says that scientists ought to use the appropriate political values — that is, the values held or endorsed by the public or its representatives — at least when those values are informed and substantively reasonable.[4] The most straightforward argument for this view grounds it in considerations of democracy or political legitimacy. If certain value choices are going to ultimately influence policy, then the public or its representatives have a right to make those choices (Douglas 2005; Intemann 2015; *cf*. Steele 2012; Kitcher 2001).

There are, of course further possibilities, and these views can be combined in more complex ways (e.g. requiring scientists to use political values in some domains, while permitting them to use their personal values in others). But if, for simplicity, we stick to these three primary

[2] In some cases, the justification for incorporating values into the scientific process dictates an answer. Feminist critiques of historically androcentric fields, for example, suggest that non-androcentric values are needed as a corrective. I set aside such cases in this paper.

[3] Mara Walli, Matthew Wong, and I discuss this at length in a work-in-progress.

[4] I set aside, then, cases where the values, say, of a policy-maker are unreasonable, in the sense that they lie outside the range of values that ought to be tolerated in a liberal society. In such cases, an advocate of the political view may permit or require scientists to reject those unreasonable values. (See e.g. Resnik 2001.) Also, in this paper I will set aside the important question of what the political view ought to say when the values of the public diverge from the values of policy-makers. The answer to this question, I think, will depend on one's theory of political representation.

options, I think the third, which I will call the *political view*, is the most attractive. More precisely, I think that in most cases where values are called for in core parts of the scientific process, scientists should privilege political values.[5] The most obvious concern with this view, and one that has received much attention from its advocates, is that it doesn't seem practical. It isn't feasible to ask citizens or policy-makers to weigh in at every point in the scientific process where values are required, and even if we could, non-experts often will not have the scientific background to fully understand the options before them. Substitutes for actual participation on the part of policy-makers or the public, such as asking scientists to predict what the public would choose or to determine what values policy-makers would hold upon reflection, seem to place unreasonable epistemic demands on scientists.

Douglas (2005), Intemann (2015), Guston (2004), and others have argued that these problems aren't insurmountable, by suggesting specific ways that the concerns of policy-makers and the public can be brought into the scientific process. And Kevin Elliott (2006; 2011) has suggested a more general way we might make progress. The political view goes hand-in-hand with a view of the relationship between science and policy that is widely-held: that the role of a scientist is to promote informed decision-making by policy-makers.[6] Bioethicists have extensively discussed how health care professionals can promote informed decision-making on the part of patients and research subjects. Theoretical and empirical research has led to a range of suggestions for how physicians can promote informed decision-making, even in cases where a patient's values may be uncertain, different research subjects may hold different values, and so

[5] This, of course, is proposed as a principle of professional ethics - not e.g. a legal requirement.

[6] See also Resnik (2001), Martin and Schinzinger (2010), and Schroeder (2016) for theoretical defenses of this idea, which is consonant with the mission statements of many scientific organizations and associations.

forth. Elliott's hope is that many of these suggestions can be adapted to the scientific case, or at least a parallel research program could be carried out, informed by the work of bioethicists.[7]

It is, of course, far from established that these proposals will work, but the range of options on the table strikes me as cause for optimism. And even if these solutions don't work in all cases, there is still bite to the political view, since it could still tell scientists to use political values *when they can determine those values*. Accordingly, in this paper I would like to describe a different and I think deeper concern with the political view, one which has been conspicuously absent from the literature thus far. In requiring scientists to guide certain aspects of their work by political values, we will sometimes in effect ask that they support political causes they may personally oppose and bar them from fully advocating for their preferred policy measures. We are, then, depriving scientists of important political rights possessed by the general public. In the remainder of this paper, I will spell out this objection more fully and explain why I think it has significant moral force. In the end, I will suggest that although there is reason to think that the objection doesn't ultimately undermine the political view, it nevertheless constitutes a significant cost that accompanies that view, which its proponents need to acknowledge.

2. Two Cases Where the Political View Seems Troublesome

The literature on values in science is vast and diverse, and so it will be useful to have some particular examples in mind. First, consider Douglas's (2000; 2009) argument that scientists should or must appeal to value judgments when resolving certain uncertainties that arise during the scientific process. Scientists conducting research into the potential carcinogen dioxin, for example, were faced with liver samples which had tumors that could not clearly be

[7] See also Schroeder (2016) for how this might go.

categorized as malignant or benign. In resolving such borderline or ambiguous cases, Douglas argues that scientists should appeal to contextual values, when the constitutive norms of science don't dictate any resolution. In this case, health-protective values would lead scientists to classify borderline samples as malignant; while concerns about overregulation would lead scientists to classify those same samples as benign (Douglas 2000).

Second, consider the many choices that scientists have to make when preparing their results for presentation. How should uncertainty be characterized? (Should 90% or 95% confidence intervals be used?) Which study results should be highlighted? (Which drug side effects should be discussed at length, and which included as part of a long list?) How should statistics be summarized? (As means or medians? Should results be broken down by gender, or presented only in aggregate?) In making choices like these, scientists frequently must appeal to values — to decide, for example, which pieces of information are important and which are not.[8]

It is, I presume, fairly uncontroversial that these value choices — how to resolve uncertainties in the research process and how to present results — can influence policy in foreseeable ways. Douglas, for example, argues that this is the case in the dioxin studies. Classifying borderline samples as malignant will make dioxin appear to be a more potent carcinogen, likely leading policy-makers to regulate it more stringently (2000, 571). Keohane, Lane, and Oppenheimer (2014) show how a presentation choice made by the Intergovernmental Panel on Climate Change led to poor policy outcomes, which likely could have been avoided by presenting information differently. More generally, we know from a wealth of studies in psychology and behavioral economics that the way information is presented to someone can strongly influence her subsequent choices (Thaler and Sunstein 2008), and there have been

[8] For discussions, see Elliott (2006), Hardwig (1994), Keohane, Lane, and Oppenheimer (2014), Resnik (2001), and Schroeder (2016).

several influential commentaries calling for scientists to more carefully "frame" their results (Nisbet and Mooney 2007; Lakoff 2010). So it seems straightforward that the value choices made by scientists can predictably affect policy.

If these value choices can influence policy, then in directing scientists to make them in accordance with political values — as opposed to the scientists' personal values — we are asking scientists to characterize policy-relevant material in a way that may promote an outcome they strongly disfavor. For example, suppose the scientists in Douglas's dioxin study value public health much more than they value keeping industry free from overregulation, but the public and its elected representatives have the opposite view. Further, suppose both views are substantively reasonable, in that they are within the range of policies eligible for adoption through democratic processes. In this case, the political view would tell the scientists to categorize borderline samples as benign, since that would better cohere with the public's values. This could make dioxin appear to have minimal carcinogenic effects, predictably leading to less regulation than would have occurred had the scientists classified borderline samples according to their own, health-protective values. Similarly, suppose an environmental economist conducting an impact study of a proposed construction project is herself deeply committed to the preservation of natural spaces. Nevertheless, if the public is strongly committed to economic development, the political view would require her to put front-and-center a detailed breakdown of the economic consequences of construction, while describing the ecological costs more briefly or in a less prominent place — likely frustrating her desire for preservation.

Notice that the concern here is not simply that scientists are being asked to provide information that will lead to an outcome they disfavor. I take it that any reasonable approach to scientific ethics will require that scientists communicate honestly, even in cases where that

promises to yield policies they don't like. Similarly, I presume that scientists must also be forbidden from presenting information in ways that, though technically accurate, are nevertheless misleading. The problem here is that Douglas's scientists are being asked to characterize results in one way (as benign) that could, *with equal scientific validity*, have been characterized differently (as malignant). And our environmental economist is being asked to present her results in one way (highlighting economic benefits), when an alternate presentation (one highlighting ecological costs) would be equally honest, accurate, objective, transparent, clear, and so forth. In each case, then, we have a collection of underlying data which can be described or characterized in different ways, neither of which appears to be more scientifically valid than the other. The political view insists that scientists choose the description grounded in values they don't accept and which seems likely to promote policy outcomes they disfavor. In this respect, the political view requires scientists to in effect advocate for, or at least tilt the playing field towards, political views they disagree with.[9]

3. Elliott and The Principle of Helpfulness

This seems clearly to be a significant imposition on scientists and thus a cost of the political view. It is therefore surprising that, so far as I can tell, philosophers who have argued for the political view have not commented on it. This is most striking in Elliott's work. Elliott, recall, argues that scientists should aim to promote informed decision-making among policy-makers, in something like the way physicians should aim to promote informed decision-making among patients. Standard accounts in bioethics say that it is the patient's values that carry the

[9] Can't we let the scientists advocate for their preferred positions in other ways? We could let scientists present their preferred interpretation separately. But if the political view is to have bite, presumably these alternate results will have to be clearly designated so and offered in a less prominent place (e.g. in an appendix or online supplement). And we should of course permit scientists to advocate for their views outside of their scientific papers/reports. But it seems likely that these (private) statements will carry much less policy weight than their scientific ones.

day: in normal cases, the physician's job is to help a patient make decisions that cohere with her own values. If the scientific cases is analogous, then the scientist's job is to help policy-makers make decisions that cohere with their (or the public's) values. This, in turn, suggests that scientists should use political values when resolving uncertainties, presenting results, and so forth. In other words, Elliott's proposal seems to imply the political view.[10]

The main defense Elliott offers for this view, however, relies on Scanlon's "Principle of Helpfulness":

> Suppose I learn, in the course of conversation with a person, that I have a piece of information that would be of great help to her because it would save her a great deal of time and effort in pursuing her life's project. It would surely be wrong of me to fail (simply out of indifference) to give her this information when there is no compelling reason not to do so.[11]

Elliott sums up the idea this way: "[I]n situations where one can significantly help another individual by engaging in an action that requires little sacrifice, it is morally unacceptable not to help" (2011, 139). If the political view, however, requires characterizing data or presenting information in ways that promote policy choices a scientist strongly opposes, then this Principle doesn't apply. When the pro-health scientist is required to classify ambiguous samples as benign, that does involve a sacrifice. A refusal to do so — which would hinder the pro-industry policy-maker's ability to make an informed regulatory decision — would not be done "simply out of indifference". It would be done out of the scientist's desire to protect public health.

---

[10] In some work, Elliott appears to suggest that transparency about values may be enough (Elliott and Resnik 2014). That is, he doesn't seem to place (many) constraints on scientists' value choices, so long as they are open about those choices. If that is Elliott's view — and it is not clear to me that it is — it strikes me as in tension with his insistence that scientists promote informed decision-making. Surely I can better help you make a decision that coheres with your values by working from your values, rather than by working from my own values (even if I am open about what I am doing). Further, even if scientists are open about their value choices, policy-makers frequently won't have the technical expertise to be able to reinterpret a scientific study, replacing one set of values (the scientist's) with another (their own). (If values could so easily be swapped out by non-specialists, then much of the debate about values would be unimportant. Transparency is all we would require.)

[11] Scanlon (1996, 224), quoted in Elliott (2011, 139).

(Similar things, obviously, can be said about the environmental economist asked to highlight the economic aspects of a proposed construction project.)

Scanlon's Principle of Helpfulness is a quite weak one, applying only in cases where the agent in question can put forward no significant burden of compliance. That Elliott uses it to justify his informed decision-making framework, and implicitly the political view, suggests that he thinks that such a view doesn't impose significant burdens on scientists. But if what I've said has been correct, that is wrong. Even if the political view is justified — and, as I've said, I think it is — we need to recognize that it asks a lot of scientists in cases where their values diverge from those of the relevant political body.

4. Physicians vs. Scientists

This, however, brings up an interesting question. If Elliott is right that the scientific case is analogous to the biomedical case, then shouldn't informed consent requirements in medicine be treated as similarly burdensome? Few bioethicists, though, would have sympathy for a physician who claimed that seeking informed consent constituted a significant ethical burden. (They may have sympathy for the claim that seeking informed consent is burdensome in more mundane ways — e.g. too time-consuming — but those complaints seem very unlike the scientists'.) I think that there is an important difference between the cases, which will help us to more clearly understand why the scientist is often burdened in a way that carries moral weight, while the physician normally is not.

We can see this by constructing a case which seems to put a physician in a position like the scientist's. Consider Jane, a doctor who strongly believes that the end of life for terminal patients is greatly enhanced by effective pain management, even if doing so shorten's the

patient's life or impairs his consciousness. For this reason, Jane has chosen palliative care as her specialty, making it her life's work to help dying patients avoid unnecessary pain. One of her patients, John, has continually insisted that he wants to remain as lucid as possible, even if that means agony. As he lies here, in agony, Jane suspects that if she framed the information properly — highlighting a medication's ability to relieve pain, while downplaying its cognitive effects — she might be able to get John to accept it. And accepting the medication, Jane strongly believes, would be much better for John. Nevertheless, standard interpretations of informed consent forbid her from doing so. Knowing that John is especially concerned about lucidity, she is ethically bound to highlight that information when informing him of his options. Unsurprisingly, John declines the pain medication and experiences what Jane regards as an awful death — precisely the kind of thing she went into palliative care to prevent.

Like our pro-health scientist, Jane has been asked to present information in a way that ultimately frustrates her deeply-valued goals. But imagine Jane complains to the ethics board at her hospital, arguing that it is burdensome to ask her to highlight to John the effects of pain medication on lucidity, because doing so would frustrate her deeply-held values. This complaint doesn't strike me as at all compelling. Why? Because Jane's values shouldn't hold any sway over John's medical choices. John has the right to reject pain medication, whatever Jane (or just about anyone else) thinks about it. Put another way, John has no obligation to take Jane's wishes into consideration, when he makes his decision. His decision is ultimately *his*.

Now, imagine our pro-health scientist complains to her ethics committee, asserting that it is burdensome to ask her to present her data in a pro-industry light, when it could with equal scientific validity be presented in a pro-health light, because doing so would frustrate her deeply-held concern for public health. Or imagine the environmental economist complaining about

having to foreground the economic benefits of the proposed construction project, since doing so will make it more likely that the project is approved and another natural space will be bulldozed. If we assume that the scientists are citizens of the society in question, then their situation is different from Jane's. As citizens in a democracy, their views should hold some sway over their government's policy choices. A government does have an obligation to take its citizens' views into consideration when making policy decisions. And when the government ultimately acts, it does so on the scientists' behalf. The decision is, in part, the scientists'.

The scientists, then, are stakeholders and even part-decision-makers in the associated policy-decisions, in a way that Jane is not a stakeholder in John's decision. This is true even if Jane cares more about John's decision than our scientists care about the policy decisions. We can see, then, that the political view isn't burdensome simply because it directs scientists to promote or advocate for outcomes they disfavor. It is burdensome because it sometimes directs scientists to promote or advocate for disfavored views, on matters that they have a right to speak on, to a body that purports to act on their behalf. This is what gives their burden its moral significance.[12]

5. Justifying the Burdens of the Political View

Some scientists have recognized the burdens that even neutrality — let alone the political view — would impose on them.

> Conservation biology is inescapably normative. Advocacy for the preservation of biodiversity is part of the scientific practice of conservation biology. If the editorial policy of or the publications in [the journal] *Conservation Biology* direct the discipline toward an "objective, value-free" approach, then they do not educate and transform society… To pretend that the acquisition of "positive knowledge" alone with avert mass extinctions is misguided… Without openly acknowledging such a perspective,

[12] What about cases where the scientists are not citizens of the society in question? In some cases, we can still make out a stakeholder claim. (When it comes to climate change, for example, we are all stakeholders in U.S. climate policy.) But such cases raise complications which I unfortunately can't discuss in a short paper like this one.

> conservation could become merely a subdiscipline of biology, intellectually and functionally sterile and incapable of averting an anthropogenic mass extinction. (Barry and Oelschlaeger 1996)[13]

Most conservation biologists enter that field because of a strong commitment to the value of biodiversity and the preservation of nature (Marris 2006). Similar things are surely true of other scientific disciplines. (My experience has been that public health researchers and economists studying inequality disproportionately share certain political values.) To the extent that these values diverge from the values of the public and its representatives, the political view would require these scientists to continually characterize their results in ways structured by a value system they find unacceptable. (In this respect, things would be quite different for, say, climate scientists. Although their work is controversial, it nevertheless is founded on values that are widely shared. The potentially catastrophic consequences of climate change are ones that virtually everyone cares about. Climate change deniers typically object to the *empirical* claims made by climate scientists - not to the basic values they hold.)

Is it fair, then, to tell a conservation biologist, who perhaps entered the field because of her love for natural spaces and has spent the bulk of her life collecting information that she hopes can be used to preserve them, that she is nevertheless ethically bound to resolve uncertainties in her research in ways favorable to economic growth, or to present her results in ways that highlight the economic value (as opposed to, say, the private or aesthetic value) of undeveloped land? I don't have a full answer to this question — such an answer would require more empirical information, as well as a fuller discussion of political philosophy — but I think we can see how the argument would go. There are a range of situations in which we impose significant

[13] This article was followed by a collection of commentaries, most of which generally supported the authors' views. Similar proposals seem to crop up frequently among conservation biologists, and are generally endorsed by those in the field (Marris 2006).

restrictions on speech and advocacy for people in important social positions. The Code of Conduct for U.S. judges, for example, bars judges from publicly endorsing candidates for political office and from making speeches for political organizations.[14] Uniformed U.S. military personnel are not permitted to participate in political fundraising, speak at political events, or display political signs, even on their private vehicles.[15] Other constraints on speech and advocacy seem ethically appropriate for politicians, police officers, lawyers, and others.

So, if there is an important public good served by constraining scientists' advocacy, it doesn't seem in principle problematic to do so. Two arguments along these lines seem promising. First, a distinctly political approach might argue that although imposing this burden on scientists does restrict important political rights of speech and advocacy, it is done in order to expand the political rights of others. By requiring scientists to work from the values of the public, the ability of the public to make informed policy choices and to effectively advocate for their own positions is enhanced. Thus, although the political view constitutes a loss of political freedom to scientists, that loss is more than balanced by the gain in political freedom to the public as a whole. (A view like this seems generally consistent with an approach to democracy like Brettschneider's (2007).)

Second, a straightforwardly consequentialist argument could point out the terrible consequences that threaten to follow if the public and/or policy-makers distrust scientific results. One of the primary arguments that has been put forward in favor of informed-consent approaches in bioethics has been that it promotes trust on the part of patients. Similarly, Elliott's informed decision-making approach — which implies the political view — seems like a promising way to

[14] http://www.uscourts.gov/judges-judgeships/code-conduct-united-states-judges

[15] http://www.dtic.mil/whs/directives/corres/pdf/134410p.pdf

promote trust in science (Elliott 2011, 133-6; *cf*. Hardwig 1994; Resnik 2001). If, then, the political view proves to be an effective way of promoting public trust in science, which in turn heads off the problems that ensue when policy-makers disregard science, that could justify imposing significant burdens on scientists.

Neither of these defenses, of course, is anywhere near complete. But both do strike me as quite reasonable, and so I don't think the concerns I've discussed in this paper should lead proponents of the political view to give up that position. That said, it is important to note the form that these defenses take. Neither attempts to show that the burden on scientists is not morally significant (as, perhaps, we might be inclined to say about the complaint of the palliative care physician). Instead, they each point to compensating benefits — not necessarily enjoyed by the scientists in question — which morally outweigh the scientists' burden. This means that the political view, even if it is justified, comes at a real cost to scientists, which is something its proponents need to acknowledge.

References

Barry, Dwight and Max Oelschlaeger. 1996. "A Science for Survival: Values and Conservation Biology," *Conservation Biology* 10: 905-11.

Brettschneider, Cory. 2007. *Democratic Rights: The Substance of Self-Government*. Princeton University Press.

Douglas Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67: 559-79.

Douglas, Heather. 2005. "Inserting the Public Into Science." In *Democratization of Expertise? Exploring Novel Forms of Scientific Advice in Political Decision-Making*, ed. Sabine Maasen and Peter Weingart, 153-169. Springer.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

Elliott, Kevin C. 2006. "An ethics of expertise based on informed consent." *Science and Engineering Ethics* 12: 637-61.

Elliott, Kevin C. 2011. *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Rsearch*. Oxford University Press.

Elliott, Kevin C. and David B. Resnik. 2014. "Science, Policy, and the Transparency of Values." *Environmental Health Perspectives* 122: 647-50.

Guston, David. 2004. "Forget Politicizing Science. Let's Democratize Science!" *Issues in Science and Technology* fall 2004.

Hardwig, John. 1994. "Toward and Ethics of Expertise." In *Professional Ethics and Social Responsibility,* ed. Wueste, 83-101. Roman and Littlefield.

Hausman, Daniel. 2015. *Valuing Health: Well-Being, Freedom, and Suffering*. Oxford University Press.

Hoffman, George and William Stempsey. 2008. "The Hormesis Concept and Risk Assessment: Are There Unique Ethical and Policy Considerations?" *BELLE Newsletter* 14: 11-17.

Intemann, K. 2015. "Distinguishing between legitimate and illegitimate values in climate modeling." *European Journal for Philosophy of Science* 5: 217-32.

Keohane, Robert O., Melissa Lane, and Michael Oppenheimer. 2014. "The ethics of scientific communication under uncertainty." *Politics, Philosophy & Economics* 13: 343-368.

Kitcher, Phillip. 2001. *Science, Truth, and Democracy*. Oxford University Press.

Lakoff, George. 2010. "Why it Matters How We Frame the Environment." *Environmental Communication* 4: 70-81.

Marris, Emma. 2006. "Should conservation biologists push policies?" *Nature* 442: 13.

Martin, Mike and Roland Schinzinger. 2010. *Introduction to Engineering Ethics (2nd ed.)*. McGraw-Hill.

Nisbet, Matthew and Chris Mooney. 2007. "Framing Science." *Science* 316: 56.

Reiss, Julian. 2013. *Philosophy of Economics: A Contemporary Introduction*. Routledge.

Resnik, David. 2001. "Ethical Dilemmas in Communicating Medical Information to the Public." *Health Policy* 55: 129-49.

Scanlon, Thomas. M. 1996. *What We Owe to Each Other*. Harvard University Press.

Schroeder, S. Andrew. 2016. "Communicating Scientific Results to Policy-Makers." Paper presented at the American Philosophical Association Conference (Pacific Division). Available at <http://apa-pacific.org/framed/download.php?file=200.pdf>.

Shrader-Frechette, Kristin. 1994. *Ethics of Scientific Research*. Rowman and Littlefield.

Shrader-Frechette, Kristin. 2008. "Ideological Toxicology: Invalid Logic, Science, Ethics About Low-Dose Pollution." *BELLE Newsletter* 14: 39-47.

Steele, Katie. 2012. "The Scientist qua Policy Advisor Makes Value Judgments." *Philosophy of Science* 79: 893-904.

Stiglitz, Joseph E., Amartya Sen, and Jean-Paul Fitoussi. 2010. *Mis-measuring Our Lives: Why GDP Doesn't Add Up*. The New Press.

Thaler, Richard and Cass Sunstein. 2008. *Nudge*. Yale University Press.

Two Roads Diverge in a Wood: Indifference to the Difference Between 'Diversity' and 'Heterogeneity' Should Be Resisted on Epistemic and Moral Grounds

Anat Kolumbus*, Ayelet Shavit* and Aaron M. Ellison

,,,
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference

from *The Road Not Taken*, by Robert Frost (1916)

Abstract:

We argue that a conceptual tension exists between "diversity" and "heterogeneity" and that glossing over their differences has practical, moral, and epistemic costs. We examine how these terms are used in ecology and the social sciences; articulate a deeper linguistic intuition; and test it with the *Corpus of Contemporary American English (COCA)*. The results reveal that 'diversity' and 'heterogeneity' have conflicting rather than interchangeable meanings: heterogeneity implies a *collective* entity that *interactively integrates* different entities, whereas diversity implies *divergence*, not integration. Consequently, striving for diversity alone may increase social injustice and reduce epistemic outcomes of academic institutions and governance structures.

* Equal main contributors.

Key words: collectivity, diversity, ecology, heterogeneity, injustice, institutional diversity.

Acknowledgments: We deeply thank the many different scholars, from very different disciplines, whose work and ideas helped us develop the ideas in this paper. In particular we want to mention Tal Israeli, Tamar Sovran, Nadav Sabar, Daryl G. Smith and Elihu Gerson. They all responded to a single email from an anonymous B.A. student with the same rigor, enthusiasm and respect as to an established full professor, and thus demonstrated the true spirit of academic inclusiveness this paper seeks to explicate. AS's work is supported by Tel Hai College and the ISF (Israeli Science Foundation) grant 960/12 and AME's work on diversity, heterogeneity, and inclusivity in science is supported by the Harvard Forest, and by grant DBI 14-59519 from the US National Science Foundation..

1. Introduction: Diversity in the Ecological and Social Sciences

The concepts of diversity and heterogeneity are two basic types of dissimilarity that are implicitly and commonly assumed to hold interchangeable meanings by scholars and laymen alike. However, when we examined their actual usage, a surprising conceptual discrepancy – in fact a tension – emerged. In this article we call attention to this tension between 'diversity' and 'heterogeneity'[1] and we argue that there are non-trivial epistemic, moral, and practical costs to science and society when this difference is glossed over.

Our critical examination is part of a large body of literature on the benefits of diversity for science and society. There exist strong epistemic (Shrader-Frechette 2002; Longino 2002; Solomon 2006b) and moral (Haraway 1979; Fricker 2007; Douglas 2009, 2015) arguments for diversity in institutions, governance structures, and ecological systems

1 In this article, we use the analytic tradition of concept notation. If quoting the concept's usage, it will appear as "X" (e.g., Fisher's "diversity" is defined as…), when explicitly mentioned as a concept it will appear as X (e.g., the concept of diversity is…), and when implicitly mentioned as a concept it will appear as 'X' (e.g., 'heterogeneity' here describes…).

("ecosystems"). For example, empirical evidence shows that diversity improves academic performance (Gurin et al. 2004; Freeman and Huang 2015; Page 2014), because diverse individuals hold different values (Longino 1990; Harding 1991), situated knowledge (Haraway 1989), socio-gender locations (Code 2006), research styles and specialities (Gerson 2013) and conflicting theoretical scaffolds (Wimsatt and Griesemer 2007). There also are costs associated with diversity, including feelings of isolation and alienation leading to reduced academic achievements of minorities (Armor 1972; Holoien 2013) and unbridgeable disagreements among researchers that disintegrate research groups (Gerson 2013; Shavit and Silver, accepted for publication).

There also are societal costs of divergence between scientists and non-scientists. Within the social realm, increased divergence from scientific worldviews may facilitate public manipulation by spreading ignorance – agnotology (Proctor and Schiebinger 2008) – and untrue and/or unjust environmental outcomes (Shrader-Frechette 2002). Within the scientific realm, divergence exempts scientists from responsibility for not assessing carefully enough social risks of generalizing their recommendations outside the laboratory, field, or model (Douglas 2009). Given the increasing science-society divergence, it is often non-experts who engage with the public – e.g., journalists teaching politicians about climate change or students teaching the underprivileged – which further widen the separation and may also silence local knowledge (Fricker 2007), e.g. by leading experienced mothers not to consider their comprehensive understanding and information as 'knowledge' compared to a young psychology student who never held a child, or depriving those living all their life near a spring to "know" their local flow rate compared to an

ecology student or governmental regulator who read published results taken at random from nearby streams (Shavit, Kolumbus and Silver, accepted for publication).

Given the fine line between the costs and benefits of constructive and destructive dissimilarities, interrogating the most basic concepts and measurements of dissimilarity seems important and timely. This paper aims for a step in that direction.

2. Definitions of Dissimilarity

Fundamental to both diversity and heterogeneity is the concept of "variance" (Fisher 1918, 1925). Briefly, measurable properties ("variables") of a group of individual entities (a "population" of cells, organisms etc.) are rarely identical. Rather, they will take on a range of values $\boldsymbol{y} = \{y_1, y_2, y_3, \ldots y_n\}$, where the value of the variable measured for the i^{th} individual is denoted y_i. When graphed as a histogram (Tukey 1977), these values are distributed, with the most frequent values clustered around the most common one and rarer values towards the edges.

The average value of the distribution of the measured variables (its expected value $E(\boldsymbol{y})$ or its mean value $\acute{y}$), equals the sum of all the individual measurements divided by the number of individuals, n: $\acute{y} = \sum_{i=1}^{i=n} \frac{y_i}{n}$. The variance, or "spread" of the distribution is the sum of the squared differences between each individual measurement and the mean:

$\sigma^2 = \sum_{i=1}^{i=n} (y_i - \acute{y})^2$. The standard error of the mean $\left(\frac{\sqrt{\sigma}^2}{n}\right)$ provides intuitive estimates of how variable the set of measurements is. Under reasonable assumptions, ≈63% of the

measurements fall within ± 1 standard error of the mean, and ≈95% fall within ± 2 standard errors of the mean.[2]

In statistics (and hence in nearly all the social and natural sciences), means and variances are characteristics of single populations (groups of measurements), but heterogeneity usually is a composite property of a group of measurements taken from more than one population. For example, the classic analysis of variance (ANOVA) developed by Fisher (1918) is used to determine if two or more populations differ in their average measured traits (e.g., height). A basic assumption of ANOVA is that the variances of the populations being compared are equal; this is referred to as "homogeneity of variance" or "homoskedasticity". In contrast, if variances are unequal (heterogeneous or heteroskedastic), mathematical transformations of the data must be done to ensure that variances are homogeneous prior to comparing populations using ANOVA.[3] Note that 'heterogeneity' here describes only the variance as a problem to overcome in order to allow a *common basis* for comparison. Throughout the rest of this article, however, the concept of heterogeneity describes entities within a collective. "Diversity", if it is used at all in statistics, refers simply to describe a collection of datasets that describe a wide range of different, often incommensurate, variables.

In contrast, diversity is used widely in ecology (e.g., McGill et al. 2015) and the social sciences (e.g., Page 2011). Unlike variance or heterogeneity, diversity is not a simple, one-dimensional predicate. McGill et al. identified at least 15 different kinds of

2 Ellison and Dennis (2010) provide a full discussion of the assumptions behind these estimates and calculation of associated confidence intervals.

3 See Gotelli and Ellison (2012) for details and another example of a "cost" of heterogeneity.

ecological diversity; differences among them reflect the number of variables or populations that are measured (one or more), the spatial scale of measurement (local or regional), and whether it is measured within or between populations. Unlike 'variance' or 'heterogeneity' – both of which are interpretable on their own – 'diversity' has little meaning to an ecologist unless it is associated with an object. For example, the concept of *alpha* diversity refers to the number of different species in a locality, the concept of *gamma* diversity to the number of different species in a region [a collection of localities], and *beta* diversity measures population change between localities.[4]

In the social sciences, Page (2011) makes similar distinctions between three kinds of diversity: (1) *variation*, or diversity within a type, referring to quantitative differences in a specific variable; (2) *diversity of types*, referring to qualitative differences between types; and (3) *diversity of composition*, or the way types are arranged. Page's variation is directly analogous to an ecologist's alpha diversity, and his diversity of types and diversity of composition are analogous to different dimensions of an ecologist's beta diversity. Most social scientists use "diversity" as a catchall phrase not attached to any particular measured process (Page, personal communication), but we suggest that more attention should be paid to the dimensions of beta diversity.

Although 'diversity' appears to be used abstractly in common parlance and is implicitly assumed to mean something very similar to 'heterogeneity', when we examined deeply rooted linguistic intuitions of certain core examples, and tested these intuitions in large databases of linguistic usage, an interesting distinction between 'diversity' and

4 Each of these can be unweighted (i.e., simple counts of different species) or weighted by their abundance or sizes (Chao et al. 2014).

'heterogeneity' was revealed, with relevance for understanding and improving civil society and its institutions.

3. A Conceptual Tension Between Diversity and Heterogeneity

Whereas scientific language may seem indecisive or vague, artistic language can be precise and revealing. For example, Robert Frost's *The Road Not Taken* beautifully highlights diverging dimensions of a difference (i.e., 'diversity'), whereas the etymology of 'heterogeneous' implies something quite the opposite: an integration of multiple other (Gr.: *hetero*) kinds (Gr. *genus*) within a single whole.

We argue that attributing heterogeneity to something (e.g., a cell, computer, etc.) implies attributing an *integration* of mutual interactions among different entities that all belong to the same *collective*, whereas attributing diversity to a collection of objects or entities entails neither interactions nor a common collective.

An examination of English idiomatic constructions reveals clear distinctions in usage of diversity and heterogeneity. We would say that the parts of a cell or a clock are heterogeneous, but not that they are diverse. In contrast, we recognize a diverse collection of wall decorations or tools. There is an apparent semantic distinction here: cells and clocks are collectives whose functioning entails the integration of a number of interacting parts, whereas walls or garages function independently of the collection of items hanging on them. In other aspects of common usage, however, many objects in daily speech, including communities, populations, or universities, are called diverse or heterogeneous interchangeably.

The *Corpus of Contemporary American English* (henceforth: COCA; Davies 2008) provides a resource with which to examine common usage of diversity and heterogeneity in more detail. COCA contains more than 520 million words of texts, including scholarly

writing, fiction and nonfiction, newspapers and spoken recordings, and has tools to conduct complex searches for occurrences of words, phrases, parts of speech, other linguistic forms, and any combination thereof. Compilations of lists of co-occurrences (i.e., all types of words [adjectives, verbs, nouns, etc.] or specific words that appear near a target word) that can be used to infer intended meanings of predicates such as diverse or heterogeneous.

Sabar (2016) used COCA to infer motivations underlying regular co-occurrences of words. By identifying partial intersection of words that regularly co-occur more than expected by chance alone, Sabar identified *communicative strategies*: the choices of specific linguistic forms that best contribute to their intended message (e.g., "look" and "carefully" form the phrase "look carefully" that calls for visual attention). Thus, the generality of a communicative strategy that is evident in a particular example is established via a quantitative prediction of a non-random co-occurrence ("look" and "carefully" occur together and in sequence more frequently than expected by chance alone, and Sabar (2016) confirmed that "look" and "see" differ in meaning as a feature of attention by showing that "look" co-occurred more frequently with words such as "notice" than did "see").

We searched COCA and the *Wikipedia Corpus* (Davies 2015) for frequencies of "diverse" and "heterogeneous" and tested our hypotheses regarding differences in meaning between them using chi-square tests for non-random frequencies. "Diverse" occurred 12-30 times more frequently than "heterogeneous" in the corpora. In line with our hypothesis, "homogeneous", "collective", "whole", "integration" and "interaction" co-occurred significantly more frequently with "heterogeneous" than with "diverse" (improved prediction by, respectively, 58, 24, 8, 11, and 11%). Antonyms of these words ("single",

"individuals", "division", "separation") showed only random patterns of co-occurrence when they co-occurred at all (see tables 1-7 in the Appendix). A possible explanation for the latter findings is that while concepts of a collective whole seem to be more explicitly related to 'heterogeneity', words and meanings of singularity are relevant to both terms (in the case of heterogeneity they could relate both a single whole or to its parts). Nonetheless, it is evident that there is empirical support for our semantic intuition regarding 'heterogeneity' as interactions among diverse entities within a collective whole, and, perhaps more importantly, the empirical lack of a collectivist meaning for 'diversity'.

The attribute of diversity does not correctly describe collective entities because its meaning and reference are much wider than the concept of heterogeneity. A heterogeneous entity may be composed physically of nothing more than diverse entities, but as a collective, it entails multiple direct and indirect interactions, and feedbacks, among these entities. All reproducing biological groups (genomes, cells, metapopulations, etc.) are heterogeneous in the collective sense. Hence, additional information that refers to internal interactive processes improves models of heterogeneous entities and systems (Wade 1978; Roughgarden, accepted for publication). Some human groups – e.g., families, football teams or kibbutzim – would best be described as heterogeneous, whereas others – e.g., people waiting to pay the cashier – would not (Shavit 2008). There may be grave costs associated with failing to identify the goals of certain human groups as diverse or heterogeneous, as the next section portrays.

4. Illustrating the Diversity-Heterogeneity Trade-Off

4.1 Moral costs

Many – perhaps most – readers of this essay would say that promoting diversity is a social good because it is a stepping-stone to heterogeneity and thus to social justice. Although we may not yet have achieved a just and heterogeneous society, we should nonetheless promote diversity as much as possible and not dwell on the semantic particularities of distinguishing the concepts of diversity from heterogeneity. We think this line of thinking is misleading, and that the continuous focus on racial, ethnic, or gender 'alpha diversity' (i.e., headcounts) and use of the results of these measurements as a sufficient basis for discourse and policy, creates a vicious circle that may hinder social change in many of our institutions, in particular in our schools, colleges, and universities.

For example, in *Brown v. Board of Education* (1954), the Supreme Court of the United States ruled that segregation of African-American and Caucasian students in schools violated the Equal Protection Clause of the U.S. Constitution. One outcome of this decision was transporting students of different racial backgrounds into different school districts ("busing") to achieve diverse, "integrated" schools. This was intended to provide equal opportunities, academic aspirations, and achievements for all students and to improve relations among different races (Armor 1972). Unfortunately, according to some of its strongest supporters, busing did not improve academic aspirations or achievements (St. John 1975), sometimes decreased them and often worsened interracial relations: "integration … enhances ideologies that promote racial segregation, and reduces opportunities for actual contact between the races." (Armor 1972, 13).

In higher education, diversification is primarily done through "affirmative action". Many scholars support affirmative action (e.g., Bowen and Bok 2000; Rothstein and Yoon

2008), but others have argued that it leads to similar or worse outcomes than would have occurred in its absence (e.g., Sander 2004; Sander and Taylor Jr. 2012). For example, between 1988 and 2007, faculty of color made up only 17% of total full-time faculty, and that there had been little change in this number since the 1980's (Turner, González, and Wood 2008). Similar findings have been reported for the number of earned PhDs (NSF 2013).

However one thinks about affirmative action, we suggest that in the interest of promoting social justice that institutions should not measure diversity alone – how many people of different backgrounds are found at a certain time and place – nor wait for it "to work its magic" and reduce injustice. Smith (2015) identifies three problems with current mechanisms for promoting diversity in higher education: (1) responding to calls to improve diversity reactively rather than proactively, often by producing an internal quantified response to an external standardized requirement; (2) failure to include people from the many interacting parts of a university – faculty, staff, students, etc. – in discussions about diversity; and (3) making diversification into a specific program rather than an integral institutional function and goal. All of these common methods of "working towards diversity" are problematic precisely because they increase diversity but reduce heterogeneity. They track and magnify difference and divergence rather than encourage and enhance mutual interaction among all different co-occurring identity groups.

A more positive approach was reported by Walton and Cohen (2011), who conducted a very brief intervention in one's sense of social belonging (SOB) to a selective, largely Caucasian, college. After three years, there was a significant increase in the GPA (grade point average) of African-American students relative to control groups. SOB is central to a

heterogeneous community as it is a psychological aspect of being a part of an integrated collective.

We suggest that a trade-off exists between tracking diversity and building heterogeneity, which may result in a vicious circle leading to blaming those afflicted with social inequality for their under-representation. Since we are better at measuring discrete variables such as grades and gender than at measuring interactions such as SOB and research cooperation, we invest more effort in creating changes we can easily track rather than those that demand more complex, "beta type", measurements (e.g., institutional SOB, type of contacts with colleagues or task composition in the lab). As a result of neither measuring these latter dynamics nor investing in their visible change, alienation and lower academic achievements may persist among minority students and scholars (Syed, Azmitia, and Cooper 2011) even while their "diversity" increases. If this processes continues, a dangerous positive feedback may emerge, where not only will one's self-image and achievements be worsened, but also his/her social identity comes out worse than before affirmative action took place.

4.2. Epistemic Benefits

Aiming for heterogeneity rather than diversity often has epistemic benefits. Human collectives – as well as individual agents – have a variety of epistemic perspectives (Shavit, Kolumbus and Silver, accepted for publication). These perspectives differ in multiple inter-related ways, involve different backgrounds and experiences, and vary in ways of perceiving, explaining, and evaluating information about the world. Perspectives direct our attention to track a wide range of phenomena, promote diverse models to explain them (Griesemer 2014) and encourage adaptive-reflection by employing "…a variety of social perspectives, often…by taking the perspective of others" (Bohman 2006, 180).

Information is distributed asymmetrically between agents, so that some of it is known in general, some exclusive to certain groups, and some idiosyncratic to specific individuals (Sunstein 2003; Andesron 2006; Solomon 2006a; Gerson 2013); lack of interaction keeps pieces of information latent.[5] Diversity alone will not ensure that information is shared and provides fewer opportunities for agents to reflect on information that they can access only through interactions with others (Longino 2002; Tollefsen 2006).

Integrative working interaction across specialties – unlike the typical diverse-one-way adoption of ideas from one disciplinary to another – "includes coordinated efforts to pose and solve new research problems that can redefine specialty boundaries" (Gerson 2013, 516), and leads to developing new specialties. Tollefsen (2006) interweaves individual and collective knowledge in a way that demonstrates the benefits of epistemic heterogeneity. She suggested a framework of splitting a group that shares a common goal (e.g., works on a related set task or problems) into sub-groups; heterogeneity is manifested on an inter-sub-group level. Each sub-group is responsible for a different task, has its own sub-goals, and devises its own strategies and solutions. Mutual interactions result when the sub-groups return to the original group setting to present their suggestions and give feedback to other sub-groups. They encounter dissenting perspectives of out-groups and are forced to consider them and examine their own perspective closely. This self-scrutiny and actual encounters with critiques by other groups reveals problems, such as inaccuracies, leaps and gaps, and uncertainties, allowing the sub-groups and the integrated collective opportunities for self-correction (Tollefsen 2006).

5 There is an on-going discussion regarding the epistemic efficacy of deliberation, which is beyond the scope of this article.

Since all sub-groups are part of a larger community that shares a common goal, they both depend on other sub-groups and are depended upon by them. This framework is heterogeneous rather than diverse as the common goal and the inter-sub-group interactions serve to integrate the group. It also maintains differences, thus reducing the danger of group cohesiveness leading to unanimity and conformism, without promoting divergence. Such a framework increases the chances of achieving accurate results and obtaining a more just process of decision-making.

5. Conclusion

Diversity is not heterogeneity, and a continued focus on the former is not increasing the latter; instead, there is often a trade-off and tension between them. We illustrated how heterogeneity can better advance academic institutions and governess structures by integrating different people, identities, perspectives, and sources of information; it facilitates interactions among them, which have constructive epistemic and moral implications. Conversely, diversity alone often leads to divergence, is insufficient to resist social injustice and it misses epistemic opportunities that result from integrative working interactions. Institutions are often unaware of the diversity-heterogeneity tension or remain indifferent to it. They invest efforts in promoting diversity while neglecting heterogeneity, thus paying the costs of the trade-off and not reaping its benefits. Tracking alpha and disregarding beta diversity maintain this trade-off and obscures it. For moral and epistemic reasons we suggest noting this conceptual and practical difference and aiming for heterogeneity.

References

Anderson, Elizabeth. 2006. "The Epistemology of Democracy." *Episteme* 3 (1-2): 8–22.

Armor, David J. 1972. "The Evidence on Busing." *Public Interest* 28:90–126.

Bohman, James. 2006. "Deliberative Democracy and the Epistemic Benefits of Diversity." *Episteme* 3 (3): 175–91.

Bowen, William G., and Derek Bok. 2000. *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions*. Princeton, NJ: Princeton University Press.

Chao, Anne, Nicholas J. Gotelli, T. C. Hsieh, Elizabeth L. Sander, K. H. Ma, Robert K. Colwell, and Aaron M. Ellison. 2014. "Rarefaction and Extrapolation with Hill Numbers: A Framework for Sampling and Estimation in Species Diversity Studies." *Ecological Monographs* 84 (1): 45–67.

Code, Lorraine. 2006. *Ecological Thinking: The Politics of Epistemic Location*. *Ecological Thinking: The Politics of Epistemic Location*. Oxford, UK: Oxford University Press.

Davies, Mark. 2008. "The Corpus of Contemporary American English: 520 Million Words, 1990-Present." Accessed February 15. http://corpus.byu.edu/coca/.

———. 2015. "The Wikipedia Corpus: 4.6 Million Articles, 1.9 Billion Words." Adapted from Wikipedia. Accessed February 15. http://corpus.byu.edu/wiki/.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press.

———. 2015. "Politics and Science: Untangling Values, Ideologies, and Reasons." *The ANNALS of the American Academy of Political and Social Science* 658 (1): 296–306.

Ellison, Aaron M., and Brian Dennis. 2010. "Paths to Statistical Fluency for Ecologist." *Frontiers in Ecology and the Environment* 8 (7): 362–70.

Fisher, Robert A. 1918. "The Correlation between Relatives on the Supposition of Medelian Inheritance." *Transactions of the Royal Society of Edinburgh* 52:399–433.

———. 1925. *Statistical Methods for Research Workers*. *Biological Monographs and Manuals*. Edinburgh: Oliver and Boyd.

Freeman, Richard B., and Wei Huang. 2015. "Collaborating with People like Me: Ethnic Co-Authorship within the US." *Journal of Labor Economics* 33 (3(S1)): S289–318.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford, NY: Oxford University Press.

Frost, Robert. 1916. *Mountain Interval*. New York, NY: Henry Holt.

Gerson, Elihu M. 2013. "Integration of Specialties: An Institutional and Organizational View." *Studies in History and Philosophy of Biological and Biomedical Sciences* 44:515–24.

Gotelli, Nicholas J., and Aaron M. Ellison. 2012. *A Primer of Ecological Statistics. 2nd Edition*. Sunderland, MA: Sinauer Associates.

Griesemer, James R. 2007. "Tracking Organic Processes: Representations and Research Styles in Classical Embryology and Genetics." In *From Embryology to Evo-Devo*, ed. Manfred D. Laubichler and Jane Maienschein, 375–433. Cambridge, MA: MIT Press.

———. 2014. "Reproduction and the Scaffolded Development of Hybrids." In *Developing Scaffolds in Evolution, and Cognition*, ed. Linnda R. Caporael, James R. Griesemer, and William C. Wimsatt, 23–55. Cambridge, MA: MIT Press.

Griesemer, James R., and Michael J. Wade. 1988. "Laboratory Models, Causal Explanations and Group Selection." *Biology and Philosophy* 3 (1): 67–96.

Gurin, Patricia, Jeffrey S. Lehman, Earl Lewis, Eric L. with Dey, Sylvia Hurtado, and Gerald Gurin. 2004. *Defending Diversity: Affirmative Action at the University of Michigan*. Ann Arbor, MI: University of Michigan Press.

Haraway, Donna. 1979. “The Biological Enterprise: Sex, Mind, and Profit from Human Engineering to Sociobiology.” *Radical History Review* 20:206–37.

———. 1989. *Primate Visions: Gender, Race, and Nature in the World of Modern Science*. New York, NY: Routledge.

Harding, Sandra. 1991. *Whose Science? Whose Knowledge? Thinking from Women’s Lives*. Ithaca, NY: Cornell University Press.

Holoien, Deborah S. 2013. “Do Differences Make a Difference? The Effects of Diversity on Learning, Intergroup Outcomes, and Civic Engagement.” University Report, The University of Princeton.

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.

———. 2002. *The Fate of Knowledge*. Princeton, NJ: Princeton University Press.

McGill, Brian J., Maria Dornelas, Nicholas J. Gotelli, and Anne E. Magurran. 2015. “Fifteen Forms of Biodiversity Trend in the Anthropocene.” *Trends in Ecology and Evolution* 30 (2): 104–13.

National Science Foundation, National Center for Science and Engineering Statistics. 2013. “Survey of Earned Doctorates, 1998–2013 [NSF Publication No. 15-304].” Accessed February 19. http://www.nsf.gov/statistics/srvydoctorates/.

Page, Scott E. 2011. *Diversity and Complexity*. Princeton, NJ: Princeton University Press.

———. 2014. “Diversity without Silos: The Confluence of the Social and Scientific Teaching of Diversity.” *Independent School Magazine* 73 (4): 27–30.

Proctor, Robert N., and Londa Schiebinger, eds. 2008. *Agnotology: The Making and Unmaking of Ignorance*. Stanford, CA: Stanford University Press.

Rothstein, Jesse, and Albert H. Yoon. 2008. “Affirmative Action in Law School Admissions: What Do Racial Preferences Do?” Working Paper 14276, National Bureau of Economic Research.

Roughgarden, Joan. *Accepted for publication*. "Model of Holobiont Population Dynamics and Evolution: A Preliminary Sketch." In *Landscapes of Collectivity in the Life Sciences*, ed. Snait Gisis, Ehud Lamm, and Ayelet Shavit. Cambridge, MA: MIT Press.

Sabar, Nadav. 2016. "A Meaning Hypothesis to Explain Speakers' Choice of the Sign Look." PhD diss., City University of New York.

Sander, Richard H. 2004. "A Systematic Analysis of Affirmative Action in American Law Schools." *Stanford Law Review* 57 (367): 367–483.

Sander, Richard, and Stuart Taylor Jr. 2012. *Mismatch: How Affirmative Action Hurts Students It's Intended to Help, and Why Universities Won't Admit It*. New York, NY: Basic Books.

Shavit, Ayelet, Anat Kolumbus, and Yael Silver. *Accepted for publication*. "Epistemic Collectives, Heterogeneity and Injustice: The Case for Town Square Academia." In *Landscapes of Collectivity in the Life Sciences*, ed. Snait Gisis, Ehud Lamm, and Ayelet Shavit. Cambridge, MA: MIT Press.

Shavit, Ayelet, and Yael Silver. *Accepted for publication*. "To Infinity and Beyond!" Inner Tensions in Global Knowledge- Infrastructures Promote Local and pro-Active 'location' Information." *Science and Technology Studies*.

Shavit, Ayelet. 2008. *One for All? Facts and Values in the Debate over the Evolution of Altruism*. Jerusalem: Magness Press, in Hebrew.

Shrader-Frechette, Kristin. 2002. *Environmental Justice: Creating Equality, Reclaiming Democracy*. Oxford, UK: Oxford University Press.

Smith, Daryl G. 2015. *Diversity's Promise for Higher Education: Making It Work. 2nd Edition*. Baltimore, MD: Johns Hopkins University Press.

Solomon, Miriam. 2006a. "Groupthink versus The Wisdom of Crowds: The Social Epistemology of Deliberation and Dissent." *The Southern Journal of Philosophy* 44 (1): 28–42.

———. 2006b. "Norms of Epistemic Diversity." *Episteme* 3 (1): 23–36.

St. John, Nancy H. 1975. *School Desegregation: Outcomes for Children*. New York, NY: Wiley.

Sunstein, Cass. 2003. *Why Societies Need Dissent*. Cambridge, MA: Harvard University Press.

Syed, Moin, Margarita Azmitia, and Catherine R. Cooper. 2011. "Identity and Academic Success among Underrepresented Ethnic Minorities: An Interdisciplinary Review and Integration." *Journal of Social Issues* 67 (3): 442–68.

Tollefsen, Deborah. 2006. "Group Deliberation, Social Cohesion, and Scientific Teamwork: Is There Room for Dissent?" *Episteme* 3 (1-2): 37–51.

Tukey, John W. 1977. *Exploratory Data Analysis*. New York, NY: Addison-Wesley.

Turner, Caroline Sotello Viernes, Juan Carlos González, and J. Luke Wood. 2008. "Faculty of Color in Academe: What 20 Years of Literature Tells Us." *Journal of Diversity in Higher Education* 1 (3): 139–68.

Wade, Michael J. 1978. "A Critical Review of the Models of Group Selection." *The Quarterly Review of Biology* 53 (2): 101–14.

Walton, Gregory M., and Geoffrey L. Cohen. 2011. "A Brief Social-Belonging Intervention Improves Academic and Health Outcomes of Minority Students." *Science* 331 (6023): 1447–51.

Wimsatt, William C., and James R. Griesemer. 2007. "Reproduction Entrenchments to Scaffold Culture: The Central Role of Development in Cultural Evolution."

In *Integrating Evolution and Development: From Theory to Practice*, ed. Roger Sansom and Robert N. Brandon, 227–324. Cambridge, MA: MIT Press.

Appendix

Table 1. Wikipedia Corpus total target words occurrences.

Diverse	**Heterogeneous**
30967	1096

Table 2. Co-occurrences of “heterogeneous”/ ”diverse” with “interaction”. Hypothesis: “heterogeneous”-“interaction” > “diverse”-“interaction”.

	Interaction present		*Interaction* absent	
	N	%	N	%
Heterogeneous	11	18	1085	7
Diverse	49	82	30918	93
Total	60	100	32003	100

P<.001

Table 3. COCA total target words occurrences.

Diverse	**Heterogeneous**
16685	1305

Table 4. Co-occurrences of “heterogeneous”/ ”diverse” with “collective”. Hypothesis: “heterogeneous”- “collective” > “diverse”- “collective”.

	Collective present		*Collective* absent	
	N	%	N	%
Heterogeneous	5	31	1300	7
Diverse	11	69	16674	93
Total	16	100	17974	100

P<.001

Table 5. Co-occurrences of “heterogeneous”/ ”diverse” with “whole”. Hypothesis: “heterogeneous”- “whole” > “diverse”- “whole”.

	Whole present		*Whole* absent	
	N	%	N	%
Heterogeneous	7	15	1298	7
Diverse	40	85	16645	93
Total	47	100	17943	100

P<.05

Table 6. Co-occurrences of “heterogeneous”/ ”diverse” with “integration”. Hypothesis: “heterogeneous”- “integration” > “diverse”- “integration”.

	Integration present		*Integration* absent	
	N	%	N	%
Heterogeneous	6	18	1299	7
Diverse	28	82	16657	93
Total	34	100	17956	100

P<.05

Table 7. Co-occurrences of “heterogeneous”/ ”diverse” with “single”. Hypothesis: “heterogeneous”- “single” < “diverse”- “single”.

	Single present		*Single* absent	
	N	%	N	%
Diverse	77	97	16608	93
Heterogeneous	2	3	1303	7
Total	79	100	17911	100

P>.05

PSA 2016: The 25th Biennial Meeting of the Philosophy of Science Association -498-

Levels of Reasons and Causal Explanation

Abstract

My starting points are the claims that explanations are answers to why-questions, and that to answer the question why some event E occurred one must provide reasons why E occurred. The idea that all explanations of events are causal then becomes the theory that the reasons why some event occurred are its causes. My main thesis in this paper is that many "counterexamples" to this theory turn on confusing two levels of reasons. We should distinguish the reasons why an event occurred ("first-level reasons") from the reasons why those reasons *are* reasons ("second-level reasons"). An example that treats a second-level reason as a first-level reason will look like a counterexample if that second-level reason is not a cause. But second-level reasons need not be first-level reasons; nor (on my theory) need they be causes. Along the way I use the distinction between levels to diagnose the appeal of, and one main flaw in, the DN model of explanation.

1 A New Causal Theory of Explanation

It is obvious that some explanations of some phenomena speak of the causes of those phenomena. Simple examples come immediately to mind: the bridge collapsed because the wind reached a certain intensity, electrons flew off the metal because light shone on it. Much more controversial is the claim that *every* explanation of why some event happened must say something about the causes of that event. What's more, not only is it controversial whether this claim is true, it is also controversial how the claim should be understood. I have a new way of understanding the idea that all explanations of events invoke causes, one that, I think, is the most natural way to understand it. I also think that the idea, understood my way, is true (with one qualification[1]), and can be defended against the repeated claim that there exist non-causal explanations.

My theory starts with the idea, which has been held by many others, that explanations are answers to why-questions.[2] A theory of explanation, then, should say what it takes for a proposition to be an answer to a why-question. Now one standard form answers to why-questions take is "P because Q": "The tide is high because the moon is overhead" answers "Why is the tide high?" But there is another form answers to why-questions can take. The other form is "A/The reason why P is that Q."[3] Now because-answers and reasons-why answers are, in some sense, equivalent. "The tide is high because the moon is overhead" and "The reason why the tide is high is that moon is overhead" in some sense convey the same information. But I think that, for theoretical purposes, it is better to focus on reasons-answers. (I argue for this claim in (Skow 2016).)

A theory built around reasons-why answers will fill in the schema

[1] See footnote 6.

[2] Among those who hold that explanations are answers to why-questions are Hempel (1965)—with some qualifications, Bromberger (1992), and Van Fraassen (1980).

[3] I ignore here the forms used to give "teleological" explanations; I extend my theory to cover teleological explanations in (Skow 2016).

1. A reason why P is that Q iff . . .

What should the claim that "explanations of events are causal" look like, if put into the form (1)? Let "P" hold the place for a sentence that describes the occurrence of an event. (I won't try to say anything useful about which sentences do this.) Here is my proposal:[4]

(T) A reason why P is that Q if and only if the fact that Q is a cause of the fact that P.[5]

The same kinds of examples that lend credence to the idea that explanations of events are causal lend credence to its translation (T) into the language of reasons. The lighting of the fuse caused the bomb to go off; sure enough, it is also true that the reason why the bomb went off is that the fuse was lit. The electron's passing through a magnetic field caused it to accelerate; sure enough, the reason why it accelerated is that it passed through a magnetic field.

On the other hand, the same examples philosophers have thought are counterexamples to the idea that explanations of events are causal also threaten to be counterexamples to (T).

A bunch of these examples, I think, are based on the same mistake. There is a distinction to be made between "levels" of reasons. The examples fail because they confuse the two levels.[6] My aim in this paper is to introduction the distinction, and show how it can be used to defuse some examples. I will look, in particular, at Elliott Sober's claim that equilibrium explanations are non-causal, and Marc Lange's claim that "distinctively mathematical" explanations are non-causal (Sober 1981, Lange 2013).

[4]There are other theories of explanation that try to capture the idea that all explanations of events are causal—for example, (Salmon 1984) and (Lewis 1986). I do not have space here to explore the differences between their theories and mine.

[5]For stylistic convenience I sometimes speak of causation as a relation between facts, and sometimes as a relation between events. I remain neutral on which, if either, of these ways of speaking gets us closer to causation's "fundamental nature."

[6]I should say that there is one kind of counterexample that I think succeeds against (T): examples of "grounding" explanations. My true view is that every reason why a given event occurred is *either* a cause *or* a ground of its occurrence. But I will ignore grounding explanation in this paper.

2 Levels of Reasons

The distinction I want to introduce is that between

- a fact R being a reason why some event E occurred—then R is a "first-level" reason; and
- a fact F being a reason why R is a reason why E occurred—then F is a "second-level" reason, a reason why something else is a reason.

Reasons on the two different levels appear in answers to different why-questions. The first-level reasons are the facts that belong in the complete answer to the question *why E occurred*. The second-level reasons, on the other hand, belong in the answer to a different why-question: the question, concerning some reason R why E occurred, of *why R is a reason why E occurred*.

It is easy to come up with examples of first-level reasons. If I strike a match and, by striking it, cause it to light, then one reason why the match lit is that I struck it. What about an example of a second-level reason? We can find one by looking for the answer to the question of why the fact that I struck the match is a reason why the match lit. One answer (there are others) is: one reason why the fact that I struck the match is a reason why the match lit is that there was oxygen in the room at the time. In general, background conditions to a cause's causing its effect are, I hold, reasons why the cause is a reason why its effect happened. (Background conditions are not, however, the only kind of second-level reason; more on this in a bit.)

3 Second-Level Reasons Need Not Be First-Level Reasons

Here is the thesis about levels of reasons that I will defend in this paper:

> A fact can be a second-level reason without being a first-level reason. A fact F can be a reason why R is a reason why E happened, without F itself being a reason why E happened.

I say that F *need not* itself be a reason why E happened; I do not say that it *cannot*. The example I gave earlier shows that sometimes F *is* also a reason why E happened. The presence of oxygen,

besides being a reason why the striking of the match is a reason why the match lit, is also itself a reason why the match lit. But it is not always like this.

Here is an example in which a second-level reason is not also a first-level reason. Jill throws a rock at a window, Joan sticks out her mitt and catches the rock, and the window remains intact. The fact that Joan stuck out her mitt is a reason why the window remained intact. There is the first-level reason. *Why* is it a reason? The reason why it is a reason is that Jill threw a rock at the window. (You can test this with a counterfactual: if Jill hadn't thrown, certainly Joan's sticking out her mitt would not have been a reason why the window remained intact. The window wouldn't have "needed" Joan's help.) But this second-level reason is not also a first-level reason: that Jill threw a rock is *not* a reason why the window remained intact.[7]

In this case, the second-level reason that is not also a first-level reason is a fact that "corresponds" to the occurrence of an event: Jill's throwing of the rock. According to my theory (T), first-level reasons why events occur all correspond to events, since they are all causes. But not all second-level reasons are like the two examples we've seen so far (Jill's throw, the presence of oxygen); not all second-level reasons correspond to events.

In fact, I hold that laws of nature are second-level reasons that are not also first-level reasons. If I drop a rock from one meter above the ground, and it hits the ground at a speed of 4.4 m/s, the fact that I dropped it from one meter up is a reason why it hit the ground at 4.4 m/s. The law relating impact speed s to drop height d, namely $s = \sqrt{2dg}$ (assuming drag is negligible and d is small), is a second-level reason: it is a reason why my dropping the rock from one meter up is a reason why the rock was going 4.4 m/s when it landed. But it is not, in my view, also a first-level reason. It is not a reason why the rock is on the ground at 4.4 m/s.

Mentioning laws of nature probably brings to mind Carl Hempel's DN model of explanation, which says (I'm sure this is familiar) that an explanation of a fact F is a conjunction of facts that (i) entail F, and (ii) essentially contains a law among its conjuncts (Hempel 1965).

[7]This is also the kind of example many take to show that causation is not transitive; see for example (Hitchcock 2001).

Hempel's theory is not framed as a theory of the reasons why facts obtain, but it is natural to interpret it as committed to the thesis that whenever there are any reasons why some fact obtains, at least one of the reasons is a law of nature. I, along with many others, reject Hempel's theory, but I have a new diagnosis of where it goes wrong. Its mistake is to take certain second-level reasons, laws of nature, to also be first-level reasons.

I asserted without argument that laws are second-level reasons; but this is a natural view to have, on certain approaches to causation. One approach to causation takes laws to be central: whenever you have a cause and effect C and E, there are some laws connecting C to E—and C is a cause of E *because of* those connecting laws.[8] But that is just to say that whenever C is a cause of E, some law is a reason why C is a cause of E. Now I hold that when some fact F is a reason why C is a cause of E, then F is also a reason why C is a reason why E happened. So it follows from this theory of causation that laws are second-level reasons. If you start here, and in addition think that second-level reasons are always also first-level reasons, you head toward the characteristic thesis of the DN model, the thesis that among the reasons why some event happens is always at least one law. But this line of thought is fallacious, because second-level reasons need not be first-level reasons; and, on my view, laws that are second-level reasons are never first-level reasons.

I admit that I have given no direct argument that laws are not first-level reasons. I'd like to put the burden on the other side: why think the are? They are certainly second-level reasons: they are certainly reasons why causes are reasons why their effects happen. But as the Joan and Jill example shows, second-level reasons are not always first-level reasons. So why think they are in the case of laws? Certainly we have a sense that laws are "explaining something"; my view captures this sense, by assigning them the role of explaining why causes explain their effects. Why isn't that enough?

[8]Hempel endorses something like this idea about causation; see (Hempel 1965: 349). It has, of course, had many other defenders.

4 How The Levels Can Get Confused

I said that the flaw in the DN model is that it mis-classifies laws, which are second-level reasons, as first-level reasons. I also sketched an argument (with a false premise) that leads to this mis-classification: "laws are second-level reasons, and second-level reasons are always first-level reasons, so laws are also first-level reasons." But I'm not saying that Hempel or anyone else ever entertained this argument explicitly. Is there anything else to be said about how and why supporters of the DN model might have come to mis-classify laws as first-level reasons?

Yes, there is. Pragmatic effects, effects of the rules of conversation on information exchange, can produce "data" that misleadingly suggest that laws are first-level reasons.

The reasons why an event happened are the parts of the answer to the question of why it happened. So if we come across a conversation in which one person asks "Why did E happen?," and another person answers this question by citing some fact F; and if that answer strikes us as correct; then we have some good evidence that F really is a reason why E happened.

Some of the evidence that laws are (first-level) reasons why events happen appears to fit this pattern (but I will argue it does not). Imagine someone walks into the room just as the rock hits the ground at 4.4 m/s, and she sees that it hit at this speed (maybe the rock fell onto a device that measures impact speeds). A curious person, she asks me why it hit the ground at 4.4 m/s. I respond,

> Well, I dropped it from one meter up, and impact speed s is related to drop height d by the law $s = \sqrt{2dg}$ (and of course $\sqrt{2 \cdot 1 \cdot 9.8} \approx 4.4$).

Haven't I answered her question? And doesn't the law that $s = \sqrt{2dg}$ appear in my answer? If so, then the law is a reason why the rock hit the ground at 4.4 m/s—isn't it?

If the answers to these questions are "yes, yes, and yes," then, at least in some cases, a law is a reason why an event occurred. It's not hard to get from this conclusion to the claim (characteristic of the DN model) that this is so in *all* cases, and that when someone answers a

why-question *without* mentioning a law, her answer is incomplete.[9]

But the answers to these questions are not "yes, yes, and yes." To explain what I think is going on I need to introduce another distinction: the distinction between a *good response* to a question and an *answer* to a question. If someone asks a question, obviously one good way to respond is to answer the question. But not every good response is an answer.

A simple example suffices to establish this. Sally asks whether Caleb is coming to the party. I know he's supposed to go to the party. I respond by saying "He's sick." This is a good response. But it is not an answer. The only two possible answers are "yes (he's coming)" and "no (he's not coming)." I didn't say either of those things.

There is a theoretical reason why we should expect there to be good responses that are not answers. The notion of an answer is a semantic one. The relation between a proposition and a question, in virtue of which that proposition is an answer to that question, is a semantic relation. But the notion of a good response is a pragmatic one. Whether a response to a question is good is a matter of what a cooperative speaker should say. In some circumstances, a cooperative speaker should respond to a question by doing something other than, or something more than, answering the question. In the simple example, I know that if I just answer the question by saying "no," then Sally will immediately ask me why he's not coming. Since I can foresee that she'll ask that, and since I know the answer to this question too, I respond to her explicit question not by answering it, but by answering the expected follow-up question. It is okay in this case not to explicitly answer the question she asked, because what I do say, my answer to the expected follow-up, conversationally implies that the answer to her explicit question is no.

I did not, however, need to be so indirect. I could have responded by answering both questions. I could have said, "no, he's sick." Here my response is good, but again it contains information that is not part of the answer to the question she explicitly asked. What keeps it from being a bad response is that the additional information is relevant to the topic of our

[9]This "incompleteness" defense is most fully developed by Railton (1981). For one thorough argument against it, see (Woodward 2003: chapter 4).

conversation; and it is relevant because, though it is not an answer to her question, it is an answer to an expected follow-up question.

I think the same thing is going on in the dropped rock example. I responded to the question by saying

> Well, I dropped it from one meter up, and impact speed s is related to drop height d by the law $s = \sqrt{2dg}$.

My response is a good one, but (as we've seen) it does not follow that every part of my response is part of an answer to the question asked. In my view, the first part of my response—"I dropped it from one meter"—is an answer to the explicit question ("why did the rock hit the ground at 4.4 m/s?"), but the second part, the law, is not; it, instead, is an answer to an unasked follow-up why-question, a follow-up question I can anticipate would be asked immediately if I only answered the explicit question. The follow-up is, of course, why is the fact that I dropped it from one meter up a reason why it hit the ground at 4.4 m/s?

In summary: it is often a good thing to include a second-level reason in a response to the question why some event happened; but the fact that this is good thing to do is compatible with that second-level reason not being a reason why that event happened.

5 Equilibrium Explanations

I now have two distinctions: that between first- and second-level reasons, and that between a good response to a why-question an answer to a why-questions. The two together provide the key to defusing many problem cases for (T), the thesis that the reasons why something happened are its causes.

Elliott Sober argued that equilibrium explanations are not causal explanations. His main example of an equilibrium explanation was R. A. Fisher's answer to the question of why the ratio of males to females in the current adult human population is very close to 1:1 (Fisher 1931). "The main idea" of Fisher's answer, Sober reports, "is that if a population ever departs from

equal numbers of males and females, there will be a reproductive advantage favoring parental pairs that overproduce the minority sex. A 1:1 ratio will be the resulting equilibrium point" (201). Parents who overproduce the minority sex are likely to have more grandchildren. So if males outnumber females in the population, the fitter trait is to be disposed to have more female children than male; being the fitter trait, this disposition should increase in frequency, with the result that the sex ratio is pushed from male-biased toward equality. The opposite happens if females outnumber males. Now Sober claims that this is not a causal explanation, since

> a causal explanation...would presumably describe some earlier state of the population and the evolutionary forces that moved the population to its present configuration...Where causal explanation shows how the event to be explained was in fact produced, equilibrium explanation shows how the event would have occurred regardless of which of a variety of causal scenarios actually transpired. (202)

In other words: Fisher's explanation does not say, for example, that the sex ratio in the year 1000 was such-and-such, and that this caused the sex ratio in the year 1100 to be such-and-such, and so on. Instead it consists of a bunch of conditional facts: for each year in the sufficiently distant past, if the sex-ratio in that year had had any "non-extreme" value (non-extreme meaning not all males or females), then the sex ratio today still would have been 1:1.

The first thing I want to say is that Sober makes a claim about what the causes of the current sex ratio are that I reject. He thinks that the only relevant causes of the fact that the sex ratio is currently 1:1 are facts of the form *the sex ratio at time T is m:n*. I'm with those who reject this claim. The fact that the sex ratio in 1000 was m:n is "too specific" to be a cause of the current sex ratio. There is a less specific fact, the fact that the percentage of males in 1000 was not 0 or 100%, that is as well placed to be the cause. The less specific fact is "better proportioned" to the effect than the more specific one; so it gets to be the cause.[10]

[10]A "proportionality requirement" on causation is defended in Yablo (1992) and Strevens (2008). The claim that examples of explanations that, like Fisher's, abstract away from the nitty-

My disagreement with Sober might not seem to help much. Isn't Fisher's explanation still a counterexample to (T)? Even if the cause of the current sex ratio is that the sex ratio in the past was never extreme, Fisher's explanation doesn't cite this cause either; his explanation instead contains a bunch of other facts, namely the conditional facts described earlier. Doesn't it follow that these conditional facts, which are not causes, are reasons why the sex ratio is 1:1, and thus that (T) is false?

I deny that those conditional facts that Fisher offers up are reasons why the sex ratio is 1:1. But I can't just say this; for when Fisher offered those facts up in response to the question of why the sex ratio of 1:1, everyone celebrated his response, they did not reject it. How can his response be something to celebrate, if it didn't answer the question?

The distinctions I introduced earlier show why. Fisher's response was something to celebrate, because it was a *good response to the question.* But it can be a good response without containing an answer; in fact that's exactly what I think is going on.

I think that the reason why the sex ratio is now 1:1 is that the sex ratio in the past was never extreme. But this is not something anyone would believe, or even be able to come to know, without an accompanying answer to the question of *why* that is the reason. So a good response to the question of why the sex ratio is now 1:1 must include an answer to the question of why the fact that the sex ratio was never extreme in the past is a reason why it is 1:1 now. And *that's* the question that the conditionals in Fisher's response constitute an answer to. Those conditional facts are second-level reasons why some other fact is a reason why the sex ratio is 1:1.

---

gritty details of the causal process that produced the event being explained count as non-causal is repeated by Batterman in, for example, (Batterman 2000: 28) and (2010: 2). Batterman assumes that abstracting away from the details takes you away from the causes; but the proportionality requirement shows that in some cases at least this is not so. Less specific facts may be better proportioned to an effect than more specific ones.

6 "Distinctively Mathematical" Explanations

Marc Lange has recently described a class of explanations that he calls distinctively mathematical explanations, and argued that they are not causal explanations (Lange 2013). My interest is not in whether his examples qualify as non-causal by his criteria, but in whether they are counterexamples to (T). Here is one of the examples:[11]

> Why did a given person [say, Jones] on a given occasion not succeed in crossing all of the bridges of Königsberg exactly once (while remaining always on land or on a bridge rather than in a boat, for instance, and while crossing any bridge completely once having begun to cross it)?...[Because] in the bridge arrangement, considered as a network, it is not the case that either every vertex or every vertex but two is touched by an even number of edges. Any successful bridge-crosser would have to enter a given vertex exactly as many times as she leaves it unless that vertex is the start or the end of her trip. So among the vertices, either none (if the trip starts and ends at the same vertex) or two could touch an odd number of edges (488-89).

Here is what Lange says about why explanations like this one not causal explanations:

> these explanations explain not by describing the world's causal structure, but roughly by revealing that the explanandum is more necessary than ordinary causal laws are (491).

There is definitely something right, and deep, in what Lange says. But I do not think that his examples are counterexamples to (T).

Let P be the property of bridge-arrangements that a bridge-arrangement has if and only if either every land-mass or every land-mass but two is met by an event number of bridges. The (supposed) answer to the question of why Jones failed that Lange presents boils down to this:

[11] This example is also discussed in detail by (Pincock 2007).

(2) The bridges of Königsberg lacked P; and, necessarily, if a bridge arrangement lacks P, then no one can cross all the bridges exactly once.[12]

Now if (2) really is the answer to the question, then my theory is false. So is (2) the answer? There are two parts to (2). First is the fact that the bridges lacked P. Now it is no problem for my theory to recognize that this fact is a reason why Jones failed. For this fact is certainly a cause of his failure. The challenge to my theory comes if the second fact in (2) is a reason why Jones failed. For the second fact, that necessarily, no one can cross all the bridges exactly once, if the bridges lack P, cannot be a cause of Jones' failure.

I want to say the same thing about this example that I've said about the others. (2), I maintain, is not an answer to the question of why Jones failed. (2) contains an answer *as a part*—the fact that the bridges lacked P. But it has another part, the necessary truth, that is not part of the answer. How is this compatible with the evident fact that (2) is a really good thing to say in response to the question of why Jones failed? Because the part of (2) that is not an answer to this question *is* an answer to an obvious follow-up why-question, namely, why is it that the bridges' lacking P is the reason why Jones failed?

Lange's diagnosis of this example, and the others he discusses, is quite sophisticated, and I don't have the space here to go in to all the things he says about them. Let me at least, however, mention one further thing he says. At one point he writes, "Even if [these examples] happen to appeal to causes, they do not appeal to them as causes...any connection they may invoke between a cause and the explanandum holds not by virtue of an ordinary contingent law of nature, but typically by mathematical necessity" (496). I am quite taken by this idea that an answer to a why-question might appeal to causes but not appeal to them *as* causes. What might this mean, in terms of reasons why? Here is a natural suggestion: maybe in some cases a cause is a reason why its effect happened, but it is false that the *reason why* the cause is a reason why its effect happened is that it is a cause. The suggestion continues: cases like that are examples

[12]I'm going to take Lange's qualifications about always remaining on land etc. as given.

of "non-causal explanations."

I think the suggestion is plausible: if there truly are cases like that, they should be counterexamples to my theory. They are not, however, counterexamples to my theory as stated. I should amend my theory to make it more vulnerable:

(T2) A reason why P is that Q if and only if (i) the fact that Q is a cause of the fact that P, and (ii) the reason why the fact that Q is a reason why P is that the fact that Q is a cause of the fact that P.

Now the question is whether the Königsberg example, or any other example, is a counterexample to (T2). I have a lot of thoughts about this, but can only be brief here. Lange's idea is that since the "connection" between the bridges' lacking P, and Jones' failure, is secured by a mathematical truth (a theorem of graph theory), the bridges' lacking P, while a reason, is not a reason because it is a cause. I reject this claim. Even if the connection is secured by a mathematical truth, the cause is still a reason because it is a cause. This assertion requires defense, but I don't have the space to defend it here.

7 Conclusion

In this paper I have presented a new causal theory of explanation that says that the reasons why an event occurred are its causes. I also drew two distinctions: that between the reasons why E happened, and the reasons why those reasons are reasons; and that between an answer to a why-question, and a good response to a why-question. I used these distinctions to defend the theory against the claim that equilibrium explanations and distinctively mathematical explanations are non-causal; and I believe the distinctions can be used to defend it against a wide variety of other examples.

References

Batterman, Robert (2000). "Multiple Realizability and Universality." *British Journal for the Philosophy of Science* 51: 115-45.

— (2010). "On the Explanatory Role of Mathematics in Empirical Science." *British Journal for Philosophy of Science* 61: 1-25.

Bromberger, Sylvain (1992). *On What We Know We Don't Know.* The University of Chicago Press and CSLI.

Fisher, R. (1931). *The Genetical Theory of Natural Selection.* Dover.

Hempel, Carl (1965). "Aspects of Scientific Explanation." In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* The Free Press, 331-496.

Hitchcock, Christopher (2001). "The Intransitivity of Causation Revealed in Equations and Graphs." *The Journal of Philosophy* 98: 273âĂŞ299.

Lange, Marc (2013). "What Makes a Scientific Explanation Distinctively Mathematical?" *British Journal for the Philosophy of Science* 64: 485-511.

Lewis, David (1986). "Causal Explanation." In *Philosophical Papers, Volume II*. Oxford University Press.

Pincock, Christopher (2007). "A Role for Mathematics in the Physical Sciences." *Nous* 41: 253-75.

Railton, Peter (1981). "Probability, Explanation, and Information." *Synthese* 48: 233âĂŞ256.

Salmon, Wesley (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton University Press.

Skow, Bradford (2016). *Reasons Why.* Oxford University Press.

Sober, Elliott (1983). "Equilibrium Explanation." *Philosophical Studies* 43: 201-10.

Strevens, Michael (2008). *Depth.* Harvard University Press.

Van Fraassen, Bas C. (1980). *The Scientific Image.* Oxford University Press.

Woodward, James (2003). *Making Things Happen.* Oxford University Press.

Yablo, Stephen (1992). "Mental Causation." *The Philosophical Review* 101: 245-80.

In Defense of the Actual Metaphysics of Race

Abstract. In a recent paper, David Ludwig (2015, 244) argues that "the new metaphysics of race" is "based on a confusion of metaphysical and normative classificatory issues." Ludwig defends his thesis by arguing that the new metaphysics of race is non-substantive according to three notions of non-substantive metaphysics from contemporary metametaphysics. However, I show that Ludwig's argument is an irrelevant critique of actual metaphysics of race. One interesting result is that actual metaphysics of race is more akin to the metaphysics done in philosophy of science than mainstream analytic metaphysics.

1. Introduction

In David Ludwig's (2015, 44) recent article "Against the New Metaphysics of Race," he argues for the provocative thesis that "the new metaphysics of race" is "based on a confusion of metaphysical and normative classificatory issues." Furthermore, to continue to engage in such a "methodologically dubious metaphysics of race" is, in Ludwig's (2015, 262) opinion, "a bad idea." Key to Ludwig's critique is that he defines "metaphysicians of race" as "committed to the ideal of one fundamental ontology of race," much like other metaphysicians engaged in mainstream analytic metaphysics (Ludwig 2015, 245). Furthermore, for Ludwig, "the new metaphysics of race" consists of disputes about "one fundamental ontology of race" (Ludwig 2015, 245). In his critique, Ludwig focuses on two debates in the new metaphysics of race.

The first is the debate about whether races exist according to the one fundamental meaning of 'race' in current, ordinary English in the United States (Ludwig 2015, 257). I'll call this *the US race debate**.[1] According to Ludwig (2015, 251, 253, 256, 260), some interlocutors

[1] The asterisk is intentional. I'm calling this debate 'the US race debate*' because I think Ludwig has changed the focus of the relevant debate. I borrow the convention of using an asterisk to flag when the meaning of a term has been changed from Joshua Glasgow (2009, 140).

in the US race debate* are Anthony Appiah, Joshua Glasgow, Michael Hardimon, Sally Haslanger, Quayshawn Spencer, and Naomi Zack.

The second debate in the new metaphysics of race is about whether humans have races according to the one fundamental meaning of 'race' in the life sciences (Ludwig 2015, 254). I will call this *the biological race debate**. Ludwig (2015, 251, 253, 259) claims that, among others, the interlocutors of the biological race debate* are Robin Andreasen, Bernard Boxill, A.W.F. Edwards, Adam Hochman, Jonathan Kaplan, Koffi Maglo, Armand Leroi, Massimo Pigliucci, Neven Sesardic, and Alan Templeton.

Ludwig defends his thesis using an argument premised on the claim that the new metaphysics of race is non-substantive according to three notions of non-substantive metaphysics from contemporary metametaphysics: one from Eli Hirsch, one inspired from Theodore Sider, and one from Ludwig himself. The relevant background here is that recent metametaphysics has been preoccupied with what constitutes a "substantive" metaphysical dispute, which, roughly, is a dispute that is *really* about metaphysics as opposed to some other topic, like how we use language (Hirsch 2005, 67).

While I agree with Ludwig that to engage in a metaphysics of race that confuses metaphysical and normative classificatory issues is a bad idea, and while I think that the new metaphysics of race (as Ludwig defines it) might be based on such a confusion, I will show that the work that *actual* metaphysicians of race are doing involves no such confusion. In other words, the point of this paper is show that Ludwig's argument is an irrelevant critique of the actual metaphysics of race.

For clarity, by 'actual metaphysicians of race', I'm talking about the same group of scholars that Ludwig is talking about in his critique, and by 'actual metaphysics of race' I'm

talking about the same body of work that Ludwig is talking about in his critique.[2] However, unlike Ludwig (2015, 245), I will not require actual metaphysicians of race or actual metaphysics of race to be "committed to the ideal of one fundamental ontology of race," even with respect to a particular linguistic context.

I will begin by clarifying Ludwig's argument and his defense of each premise. Second, I will show that even if Ludwig's argument is a good critique of the new metaphysics of race, it's irrelevant to the actual metaphysics of race. Finally, I will provide closing remarks where, among other things, I will clarify how the actual metaphysics of race is more akin to the metaphysics done in the philosophy of science than mainstream analytic metaphysics. As for objections, I will respond to them along the way.

2. Ludwig's Argument and Its Defense

2.1 The Basic Argument

Though Ludwig does not state his argument explicitly, a charitable reconstruction of it is below:

(1) If the new metaphysics of race is non-substantive, then it is based on a confusion of metaphysical and normative classificatory issues.

(2) The new metaphysics of race is non-substantive.

(3) So, the new metaphysics of race is based on a confusion of metaphysical and normative classificatory issues.

[2] For instance, like Ludwig (2015, 244), I consider Joshua Glasgow to be an actual metaphysician of race, and, like Ludwig (2015, 263), I consider Glasgow's actual metaphysics of race to consist of work like his book *A Theory of Race* and his article "On the New Biology of Race."

Ludwig states (3) as his thesis in the first paragraph of his opening remarks.[3] Ludwig states (2) in his opening remarks as well and at several points throughout his paper.[4] Ludwig also treats (2) as a reason for adopting (3).[5] However, since there is a logical gap between (2) and (3), it's charitable to add (1) as a suppressed premise.[6]

2.2 Ludwig's Defense of His Premises

Though Ludwig takes the truth of (1) for granted, he offers three, in-depth defenses of (2) that utilize three different notions of non-substantive metaphysics. Ludwig's first defense of (2) is the following:

(4) The new metaphysics of race is substantive only if there is exactly one allowable and fundamental ontology of race for each of its race debates.

(5) If there is a plurality of legitimate biological subdivisions below the species level or a plurality of equally allowable specifications of 'race' for each race debate in the new metaphysics of race, then there is a plurality of allowable ontologies of race for each race debate in the new metaphysics of race.

(6) The antecedent of (5) is true.

(7) So, it's not the case that the new metaphysics of race is substantive.

Ludwig claims (4) in section 3.1 and justifies his constraint on substantive metaphysics from how he defines 'a metaphysics of *x*.' For Ludwig (2015, 245, 251), a project on the

[3] See Ludwig (2015, 244).

[4] See Ludwig (2015, 245, 260-262).

[5] See, especially, sections 3.1-3.3 and 4 in Ludwig (2015).

[6] [removed for blind review]

"metaphysics of *x*" assumes that metaphysicians of *x* are committed to "one fundamental ontology" of *x* that rules out "a plurality of equally allowable ontologies" of *x,* at least for the relevant linguistic context.[7] Since a substantive metaphysics of *x* must at least be a metaphysics of *x,* it follows that a substantive metaphysics of *x* requires exactly one allowable and fundamental ontology of *x.* Substituting 'race' for '*x*' gives us (4).

As for (5), Ludwig states that the first disjunct of (5)'s antecedent leads to (5)'s consequent in section 2. Here Ludwig (2015, 247) follows Kaplan and Winther (2013) in arguing that if there is a plurality of equally legitimate but distinct ways of subdividing species into "legitimate biological kinds," then "[e]mpirical evidence underdetermines the ontological status of race," which in turn, permits a plurality of allowable ontologies of race (Ludwig 2015, 246-247). In particular, Ludwig (2015, 245, 247-249) argues that "both racial realism and antirealism" are allowable ontologies of race given different equally legitimate ways of subdividing a species, and even in the same race debate. An example is how Zack (2002) uses the fact that humans have no subspecies to defend racial anti-realism in the US race debate*, while Spencer (2014) uses the fact that humans have a population subdivision that matches the current US census racial scheme to defend racial realism in the same race debate.

Ludwig states that the second disjunct of (5)'s antecedent leads to (5)'s consequent in section 3.1. In his words, "If there is a plurality of equally allowable specifications of 'race', there is also a plurality of equally allowable ontologies of race" (Ludwig 2015, 251). Interestingly, Ludwig never defends this assertion because he takes it to be obviously true.

[7] See Ludwig (2015, 251) for (4) and see Ludwig (2015, 245) for Ludwig's view on the metaphysics of *x.*

Next, Ludwig defends (6) by defending the truth of each disjunct in the antecedent of (5). As for the first disjunct, Ludwig (2015, 246-247) argues that there is a plurality of legitimate biological divisions below the species level (e.g. population subdivisions, monophyletic levels, subspecies, etc.) because, first, legitimate biological kinds are *interest dependent,* and, second, there is a plurality of "explanatory interests" among biologists in different research contexts (e.g. population genetics, phylogenetic systematics, etc.). As for the second disjunct, Ludwig reaches it by making an induction from what's going on in the two most popular race debates in the new metaphysics of race: which are the US race debate* and the biological race debate*.

Ludwig (2015, 254) argues that there is a plurality of equally allowable specifications of 'race' in the biological race debate* since biologists in different research programs use 'race' in different ways that suit their needs. For instance, Ludwig (2015, 254) points out that 'race' is often used as a synonym for 'subspecies' in systematic biology, but often used as a synonym for 'ecotype' in ecology. As for the US race debate*, Ludwig takes a more circuitous route to the conclusion that there is a plurality of equally allowable specifications of 'race' in that debate. First, Ludwig (2015, 255) appeals to Glasgow et al.'s (2009) empirical research on how Americans use 'race' to argue that 'race' is "polysemous" in the current US. Next, Ludwig (2015, 257-258) argues that the *context* for the US race debate* has not been "sufficiently specified" to narrow the debate to "exactly one fundamental candidate meaning of 'race' in the United States." Hence, according to Ludwig, from induction, the second disjunct of (6) holds as well.

Ludwig's second defense of (2) utilizes Hirsch's notion of non-substantive metaphysics. The second defense is below:

(8) A dispute is merely verbal if each side can plausibly interpret the other

side as speaking a language in which the latter's asserted sentences are true.

(9) A dispute is non-substantive if it is merely verbal.

(10) Each side can plausibly interpret the other side as speaking a language in which the latter's asserted sentences are true in the new metaphysics of race.

(11) Thus, the new metaphysics of race is non-substantive.

(8) is a direct quote from Ludwig (2015, 259), which is itself a summary of Hirsch's (2005; 2008) view on non-substantive metaphysics.

Hirsch defends his distinction between merely verbal disputes and ones that aren't with several examples from the history of science and philosophy. For instance, Hirsch (2005, 73) shows that the dispute among classical physicists about whether a projectile's final velocity is equal to its initial velocity on Earth was not a merely verbal dispute because physicists on both sides could not charitably interpret the other side's assertions as true. In other words, both sides were using the same meanings of 'projectile', 'velocity', 'Earth', etc., and what they disagreed about were the laws of motion. In contrast, Hirsch (2008, 407-408) shows that the dispute between John Locke and Joseph Butler about whether a tree can survive a change in its parts was merely verbal since either side could charitably interpret the other side's assertions as true using the other's meaning of 'identity'. In short, a merely verbal dispute for Hirsch is one where the disputants are either talking past one another or merely arguing about how we do (or should) use language.

As for (9), we can infer that it's a premise from how Ludwig (2015, 259-260) uses 'merely verbal' and 'nonsubstantive' at this point in his paper. Furthermore, Ludwig's

vocabulary here is uncontroversial since it's the same vocabulary that Hirsch (2005, 67) uses. As for (10), Ludwig endorses it when he says the following:

> Realists like Andreasen, Edwards, Leroi, Sesardic, and Spencer can interpret antirealists as speaking the truth in a language in which 'race' refers to subspecies, populations with visible traits that mark relevant biological differences, populations with cognitive differences, and so on. Antirealists like Glasgow, Lewontin, Hochman, Maglo, and Zack can interpret realists as speaking the truth in a language in which 'race' refers to genetic clusters, patterns of mating, clades, and so on (Ludwig 2015, 259-260).

Finally, Ludwig defends (2) in a third way using his interpretation of Sider's notion of non-substantive metaphysics. Ludwig's third defense of (2) is below:

(12) A dispute about an expression *E* is non-substantive if its disputants are endorsing multiple, equally joint-carving candidate meanings for *E.*

(13) The new metaphysics of race is a dispute that is non-substantive according to (12).

(14) The new metaphysics of race is non-substantive.

(12) is directly from Ludwig (2015, 261), and is a rough summary of Sider's (2011, 46-49) view of non-substantive metaphysics. Sider defends the non-joint-carving condition in his definition of 'non-substantivity' from his stipulation of what metaphysics is about.

For Sider (2011, vii) the "central task" of metaphysics is "to discern the ultimate or fundamental reality underlying the appearances." We are supposed to describe this reality using a privileged language, so-called Ontologese, which is privileged exactly because all of its expressions (e.g. terms, quantifiers, etc.) are "joint-carving," which means that they carve out the

world's fundamental structure (Sider 2011, vii).[8] So, naturally, when we find that one or more of the expressions that we've used to formulate a question *Q* does not have exactly one, best joint-carving meaning, it's likely that a debate about *Q* is not about the fundamental structure of the world, and thus, is not a substantive metaphysical debate in Sider's sense.

With that said, it's important to note that Ludwig's summary of Sider is rough, and does not reflect Sider's (2011, 49) "revised" definition of a non-substantive dispute. What Ludwig presents is Sider's unrefined view, which occurs at the beginning of section 4.2 in chapter 4 of Sider's *Writing the Book of the World*. However, later on in section 4.2, after Sider considers multiple problems with his unrefined view, he settles on what he calls his "revised" definition.[9] Nevertheless, since Ludwig uses Sider's unrefined notion of non-substantivity in his critique, that's what I'll focus on as well. However, for clarity, I'll say that (12) expresses *Sider-style non-substantivity* as opposed to Siderian non-substantivity.

In any case, Ludwig (2015, 261) asserts and defends (13) when he says that Spencer's, Leroi's, Pigliucci's, and Hochman's biological definitions of 'race' are all "equally joint-carving candidates" for 'race' because they are all "objective ways of distinguishing between populations below the species level." Furthermore, Ludwig (2015, 261-262) bolsters his support for (13) when he says that Hardimon's, Glasgow's, Feldman and Lewontin's, and Appiah's biological definitions of 'race' are also equally joint-carving candidates for 'race' because they are all "non-joint-carving" meanings.

3. Why Ludwig's Argument is an Irrelevant Critique of Actual Metaphysics of Race

[8] For Sider's clarification of "Ontologese," see Sider (2011, 171-173).

[9] For Sider's "revised" definition, see Sider (2011, 49).

Even though Ludwig has provided a valid argument that may be sound as well, it turns out that Ludwig's critique does nothing to undermine the actual metaphysics of race. The latter is partially because Ludwig's critique is not *about* the actual metaphysics of race, it's about a hypothetical metaphysics that he calls 'the new metaphysics of race'.

Remember that the new metaphysics of race is, according to Ludwig (2015, 245), and by definition, constituted by disputes about "one fundamental ontology of race." Furthermore, remember that Ludwig claims that people like Glasgow, Haslanger, Appiah, and Spencer are engaged in one such dispute, the US race debate*, and people like Andreasen, Pigliucci, Kaplan, and Templeton are engaged in another such dispute, the biological race debate*. However, these last two claims are simply false.

For one, the term 'fundamental ontology' is not even a phrase used in actual metaphysics of race. For instance, it does not appear *once* among the actual metaphysics of race that Ludwig (2015, 263-265) cites, and he cites 40 such publications. Second, some actual metaphysicians of race embrace a pluralist ontology for the nature of race in the relevant context. For example, at the beginning of Spencer's (2014, 1026) article on the "national" meaning of 'race' in the US, he concedes that ordinary Americans are using multiple "geographic" and "ethnic" meanings of 'race'. In fact, Spencer (2014, 1026) explicitly says, "Hence, I acknowledge upfront that there are several ways that Americans use 'race'."

However, Ludwig could object here. Specifically, Ludwig (2015, 257) interprets Spencer's focus on the national meaning of 'race' in the US as an endorsement of it being "the only relevant candidate meaning for philosophical debates about the referent of "race" in the United States." While the latter is a possible interpretation of Spencer's project, it's not the most charitable one given how he presents his project at the beginning of his article. Spencer (2014,

1025) begins by saying upfront that his project is merely "to debunk" the idea that "folk racial classification has no biological basis." Spencer attempts to accomplish that goal by showing that 'race', in its national meaning in the current US, is a directly referring term for a biological entity—a set of particular human populations—that presently happens to be biologically real in virtue of being a level of human population structure. Thus, given how Spencer (2014, 1026) presents his own project, his race theory is compatible with there being a pluralist nature of race in the current US context. Furthermore, this interpretation best explains why Spencer (2014, 1026) says that "there are several ways that Americans use 'race'."

There are other actual metaphysicians of race who embrace pluralism about the nature of race as well. For instance, Pigliucci and Kaplan (2003, 1162-1163) are happy to grant that both the ecotype and the subspecies are equally legitimate ways of dividing a species into biological races. It's just that they believe that humans have ecotypes, but not subspecies. In fact, Pigliucci and Kaplan (2003, 1163) explicitly say, "Races, then, can be defined and picked out in a number of ways."

Finally, there are plenty of actual metaphysicians of race who do not embrace pluralism about the nature of race, but who do entertain pluralism as a metaphysical possibility, which is enough to show that they do not presuppose that there is a single fundamental ontology of race in the relevant context. For instance, after obtaining messy results about how ordinary Americans use 'race' and race terms in a widely distributed survey, Glasgow (2009, 75) entertains the possibility that ordinary Americans are sometimes "talking past each other" when they use 'race', much like we sometimes do when we use 'jade'. In fact, Glasgow (2009, 75) explicitly says, "So maybe 'race' is used in some contexts to refer to a social kind of thing and in other contexts to a biological kind of thing." That doesn't sound like somebody who presupposes that

there is a single fundamental ontology of race in the US context. Now, even though Ludwig's argument is not about actual metaphysics of race, it could still be a relevant critique of actual metaphysics of race. So to that I now turn.

In order to know whether Ludwig's argument succeeds in critiquing the actual metaphysics of race, we need to know more about the debates among actual metaphysicians of race. Clearly, the US race debate* and the biological race debate* are not debates among actual metaphysicians of race. However, the US race debate and the biological race debate are. *The US race debate* is the debate about the nature and reality of race according to what 'race' means in the ordinary discourse of contemporary Americans, but only when 'race' is used to classify humans. The latter debate actually exists because all of the individuals that Ludwig places in the US race debate* have expressed an interest in the focus I've just articulated.[10] *The biological race debate* is the debate about whether humans have any races in a nontrivial biological sense of 'race'. The latter debate actually exists as well.[11] These are the two race debates that Ludwig was attempting to critique, and given these distinctions, we can see that Ludwig's argument really isn't relevant to these two debates.

For one, neither the US race debate nor the biological race debate satisfies Hirsch's criterion for a non-substantive dispute. The US race debate is not a merely verbal dispute because racial realists in that debate, such as Haslanger and Spencer, cannot plausibly interpret racial anti-realists in that debate, such as Appiah and Glasgow, as speaking a language in which

[10] For evidence, see Appiah (1996, 42), Glasgow (2009, 15), Haslanger (2012, 133), and Spencer (2014, 1025).

[11] For evidence, see Andreasen (1998, 200-201, 205), Pigliucci and Kaplan (2003, 1161-1164), Maglo (2011, 362-363), and Templeton (2013, 262-263).

anti-realist race theories are true, and vice versa. For instance, if Glasgow (2009, 33) is correct about (H1*) being part of the non-negotiable semantic content of 'race' in the ordinary discourse of Americans, then Spencer (2014, 1026) is incorrect about 'race' directly referring to a set of human populations in the national racial discourse of Americans, and vice versa.[12] The biological race debate is not a merely verbal dispute either. For instance, if Pigliucci and Kaplan (2003, 1165) are correct that humans subdivide into "biologically significant" ecotypes, then Hochman (2013, 347) is incorrect that humans do not subdivide into "meaningful biological units," and vice versa.

Next, even if the US race debate or the biological race debate is non-substantive in a Ludwigian or Sider-style sense, that fact does not imply a "confusion about metaphysical and normative classificatory issues" as (1) claims. This is because actual metaphysicians of race are adopting a different view of *substantive* metaphysics—namely, one that does not require metaphysical disputes about race to presuppose a single fundamental ontology of race or anything about joint-carving. Thus, while Ludwig's argument is relevant to the hypothetical new metaphysics of race, it doesn't make contact with actual metaphysics of race.

Interestingly, when Ludwig defines 'the new metaphysics of race', he anticipates the worry that his focus on it may mischaracterize actual metaphysics of race. In response, Ludwig (2015, 245) says, "However, I do not want to engage in a verbal dispute about the meaning of "metaphysics of race"… this article only challenges a certain type of metaphysics of race while proposing an alternative deflationist and normative metaphysics of race." However, this reply is

[12] (H1*) is the claim that a race is, at least, a group of human beings that is distinguished from other groups of human beings by visible physical features, of the relevant kind, that the group has to some significantly disproportionate extent (Glasgow 2009, 33).

perplexing because if the new metaphysics of race is a purely hypothetical metaphysics that does not describe the disputes in actual metaphysics of race (as I've shown), and, in addition, if the disputes in actual metaphysics of race already do away with monist and fundamentalist assumptions about race (as I've shown), it's hard to imagine what the purpose is for lodging Ludwig's critique in the first place. In any case, we can rest assured that actual metaphysicians of race are immune to Ludwig's critique because they've already been vaccinated against monist and fundamentalist assumptions about race.

5. Closing Remarks

In this paper, I've shown that Ludwig's critique of the new metaphysics of race is irrelevant to the actual metaphysics of race. However, I've said little about the conditions of substantivity that actual metaphysicians of race adopt. In addition to the bare minimum of "not talking past one another" (Glasgow 2009, 28), actual metaphysicians of race embrace disputes about how certain linguistic communities actually use 'race' (e.g. Pigliucci and Kaplan 2003, 1162-1163; Glasgow 2009, 6), and embrace disputes about how certain linguistic communities should use 'race' (e.g. Haslanger 2012, 221-247; Hochman 2014, 80). However, actual metaphysicians of race do not embrace disputes that have unimportant social and scientific consequences. For instance, Haslanger (2012, 300) motivates the US race debate by pointing out that engaging in it will help us frame and evaluate social policies and appropriately address stubborn inequalities in health. Also, Pigliucci and Kaplan (2003, 1170) point out that engaging in the biological race debate can help biologists debunk hereditarian hypotheses about race and intelligence, yield insights into human evolutionary history, and yield insights into human migration history.

Interestingly, the criteria for substantive metaphysics that actual metaphysicians of race adopt make the metaphysical disputes in the actual metaphysics of race more akin to metaphysical disputes in the philosophy of science (e.g. the species debate, the nature of natural kinds, the ontic structural realism debate, etc.) than those in mainstream analytic metaphysics (e.g. debates about the nature of fundamentality, grounding, modality, substantivity, etc.). For instance, Matthew Slater's (2015) stable property cluster theory of natural kinds has a real shot at explaining why some kinds support epistemically reliable inductions in a domain while others don't, which could help systematic biologists achieve more agreement about how they should classify organisms into species and higher taxa. So, much like disputes in the actual metaphysics of race, there are practical payoffs to science or society for engaging in metaphysical disputes in the philosophy of science. However, mainstream analytic metaphysics does not guarantee a payoff for science or society. For instance, what exactly is the payoff for science or society in debating about "the" nature of substantive metaphysics?

Perhaps Sider (2011, 47) sums up my point best when he says, "… this concept is not intended to apply to everything that might justly be called "nonsubstantive". For example, it isn't meant to apply to equivocations between distinct lexical meanings (as in a dispute over whether geese live by "the bank", in which one disputant means river bank and the other means financial bank)… Nor is it meant to capture the shallowness of inquiry into whether the number of electrons in the entire universe is even or odd (an inquiry that is substantive in my sense, but pointless)."

References

Andreasen, R. O. (1998). A New Perspective on the Race Debate. *The British Journal for the Philosophy of Science, 49*(2), 199-225.

Appiah, K. A. (1996). Race, Culture, Identity, Misunderstood Connections. In K. A. Gutmann, *Color Conscious* (pp. 30-105). Princeton: Princeton University Press.

Glasgow, J. (2009). *A Theory of Race.* New York: Routledge.

Glasgow, J., Shulman, J., & Covarrubias, E. (2009). The Ordinary Conception of Race in the United States and Its Relation to Racial Attitudes: A New Approach. *Journal of Cognition and Culture, 9*, 15-38.

Haslanger, S. (2012). *Resisting Reality.* Oxford: Oxford University Press.

Hirsch, E. (2005). Physical-Object Ontology, Verbal Disputes, and Common Sense. *Philosophy and Phenomenological Research, 70*(1), 67-97.

Hochman, A. (2013). Against the New Racial Naturalism. *The Journal of Philosophy, CX*(6), 331-351.

Hochman, A. (2014). Unnaturalised racial naturalism. *Studies in History and Philosophy of Biological and Biomedical Sciences, 46*, 79-87.

Kaplan, J., & Winther, R. (2013). Prisoners of Abstraction? The Theory and Measure of Genetic Variation, and the Very Concept of "Race". *Biological Theory, 7*(1), 401-412.

Ludwig, D. (2015). Against the New Metaphysics of Race. *Philosophy of Science, 82*(2), 244-265.

Maglo, K. (2011). The Case against Biological Realism about Race: From Darwin to the Post-Genomic Era. *Perspectives on Science, 19*(4), 361-390.

Pigliucci, M., & Kaplan, J. (2003). On the Concept of Biological Race and Its Applicability to Humans. *Philosophy of Science, 70*(5), 1161-1172.

Sider, T. (2011). *Writing the Book of the World.* Oxford: Oxford University Press.

Slater, M. (2015). Natural Kindness. *The British Journal for the Philosophy of Science, 66*(2), 375-411.

Spencer, Q. (2014). A Radical Solution to the Race Problem. *Philosophy of Science, 81*(5), 1025-1038.

Templeton, A. R. (2013). Biological Races in Humans. *Studies in History and Philosophy of Biological and Biomedical Sciences, 44*(3), 262-271.

Zack, N. (2002). *Philosophy of Science and Race.* New York: Routledge.

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

Measurement of Statistical Evidence: Picking Up Where Hacking (et al.) Left Off

Abstract Hacking's (1965) Law of Likelihood says – paraphrasing– that data support hypothesis H_1 over hypothesis H_2 whenever the likelihood ratio (LR) for H_1 over H_2 exceeds 1. But Hacking (1972) noted a seemingly fatal flaw in the LR itself: it cannot be interpreted as the degree of "evidential significance" across applications. I agree with Hacking about the problem, but I don't believe the condition is incurable. I argue here that the LR *can* be properly calibrated with respect to the underlying evidence, and I sketch the rudiments of a methodology for so doing.

Introduction

The "likelihoodist," or "evidentialist," school of thought in statistics is well known among philosophers, more so perhaps than among scientists or even statisticians, in large part due to Hacking (1965). One way to distinguish evidentialism from the other major schools – frequentism and Bayesianism – is to note that evidentialism alone focuses on the assessment of statistical evidence as its principal task, rather than decision-making or the rank-ordering of beliefs.[1]

[1] Hacking himself generally prefers the term "support" over "evidence," as does Edwards (1992), but other representatives of this school (Good 1950; Barnard 1949; Royall 1997) refer to an equivalent concept as "evidence." I prefer "evidence," since this is the familiar, albeit vague, word for what we are trying to illuminate; and I prefer "evidentialist" over "likelihoodist" as the name of the school, since the former highlights a key distinction

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

It might be thought, therefore, that evidentialism would be the predominant approach to statistical inference in science, where quantifying evidence is usually the main objective. (If you don't agree, try getting scientists to stop using the p-value as a measure of the strength of the evidence!) But frequentism, and to a lesser extent Bayesianism, predominate in the scientific literature, while evidentialism is virtually unseen. Why is this? I'm going to argue here that the fault lies with evidentialism's failure thus far to address the problem of calibrating the units in which evidence is to be measured. Since meaningful calibration is the sine qua non of scientific measurement, this turns out to be the loose thread that causes the cloth to unravel when we pull on it.

Before proceeding it may be worth noting some things I will and will not be talking about. First, I am concerned only with *statistical* evidence, and will not be considering the concept of evidence as it appears in other contexts, e.g., in legal proceedings. Second, I will treat statistical evidence as a *relationship* between data and hypotheses under a model that can be expressed in the form of a likelihood (as defined below). On this view, data do not possess inherent evidential meaning on their own, but only take on meaning in the context of their relationships to particular hypotheses, with the nature of those relationships governed by the form of the likelihood. I will not be concerned here with measurement problems associated

---

between this school and the others. By contrast, likelihood features prominently in all modern statistical frameworks.

with the data themselves.[2] Third, I am interested here solely in addressing the question of whether this relationship between data and hypotheses can be rigorously quantified. If the answer is yes, then presumably the degree of evidence could play a role in decision making (deciding how strong is strong enough when it comes to evidence) or in guiding belief, but I will not be addressing these topics here. It is one hallmark of evidentialist reasoning that statistical evidence is treated independently of these matters.

The remainder of the paper is organized as follows. In section (1) I articulate the central evidence calibration problem (ECP), and suggest reframing it in measurement terms. In section (2), I consider ways in which evidentialism's preoccupation with so-called "simple" hypotheses (as defined below) has constricted the theory, masking the true nature of the underlying measurement problem, and also obscuring the solution. In section (3) I illustrate a methodology for beginning to address the ECP once the restriction to simple hypotheses is relaxed. In section (4) I briefly consider what changes would be required to axiomatic foundations in order to accommodate this methodology while remaining true to the spirit of evidentialism's original motivating arguments.

(1) The Evidence Calibration Problem (ECP)

At the heart of evidentialism is Hacking's (1965) familiar Law of Likelihood, which says in essence that data support one statistical hypothesis H_1 over another hypothesis H_2

[2] In common usage "evidence" is often used to refer to what I am calling *data*, but "evidence" also has this other sense of being a *relationship* between data and hypotheses. In order to maintain this distinction, I will call the data "data" and the relationship "evidence."

whenever the likelihood ratio (LR) for H_1 over H_2 exceeds 1. But Hacking (1972) pointed out a problem in assigning any particular interpretation to the magnitude of the LR. In his review of Edwards (1992, orig. 1972), he says:

> "Now suppose the actual log-likelihood ratio between the two hypotheses is *r*, and suppose this is also the ratio between two other hypotheses, in a quite different model, with some evidence altogether unrelated to [the original data]. I know of no compelling argument that the ratio *r* 'means the same' in these two contexts."[3] (p. 136)

Thus we can say that, for one experiment, data support hypothesis H_1 over hypothesis H_2 with LR = 2, and, for another experiment, that a different set of data support H_3 over H_4 with LR = 20; but we cannot saying anything definite about how much more the second set of data supports H_3 over H_4 relative to the amount by which the first set supports H_1 over H_2. Edwards was well aware of this problem, saying expressly that "we shall not be attempting to make an absolute comparison of *different* hypotheses on *different* data." (p. 10). But Hacking's point cuts deep. *If the numerical value of the LR cannot be meaningfully compared across applications, in what sense is it meaningful in any one application*?

[3] Here Hacking is using "evidence" in the sense of what I am calling *data*; however, he goes on to describe what he has in mind in terms of levels of "evidential significance." He refers to the *log* LR as this is the form preferred by Edwards. Note that Hacking already appears to have been alluding to this problem in Hacking (1965), vide p. 61.

Veronica J Vieland
Philosophy of Science Assoc Biennial Meeting 2016

Hacking's criticism points to a fundamental problem for evidentialists, who appear to be able to say *whether* given data support H_1 over H_2, but not by *how much* they support H_1.[4] This is on the face of it metaphysically perplexing, but also, it leaves a gap between *support*, as Hacking's Law defines it, and a truly quantitative *weight of evidence*, which would be far more useful scientifically if only we could work out how to evaluate it.

Following the core arguments in Barnard (1949), Hacking (1965) and Edwards (1992), I will assume that the LR is the key quantity in any cogent theory of statistical evidence. But the Law of Likelihood is more specific than this assumption: it assigns a particular importance to one very narrowly conceived *aspect* of the LR, a fact that is obscured by evidentialism's focus on simple hypotheses, to which I turn next.

Before doing so, I note that resolving Hacking's problem requires unpacking his phrase 'means the same'. I think that this must be understood as 'means the same with respect to the underlying evidence,' a locution that lands us solidly in *measurement* territory. We must be able to think in terms of the underlying evidence, as something we can – at least in the abstract – conceive of independently of how we measure it. The question then becomes: How do we establish meaningful measurement units for evidence, so that a given measurement value always 'means the same' *with respect to the evidence*? This is the ECP.

And here, in a nutshell, is the evidentialist's difficulty in addressing the ECP. The LR for a simple hypothesis comparison (see below) is a single number, thus, the evidentialist is lured

[4] Royall (1997) is the only one as far as I know who argues that the magnitude of the LR *does* express strength of evidence in a comparable manner across applications. But I think his arguments on this point fail for reasons articulated in Forster & Sober (2004).

into the claim that "the LR *is* the evidence." To see the danger here, consider a mercury thermometer reading 80°F. We might say, "the temperature is 80°," but this is a circumlocution for "80 is the numerical value we assign, on the Fahrenheit scale, to the underlying temperature." Now suppose that rather than degrees, only units of volume *V* are annotated on the sides of the glass. We might be tempted to say "*V* is the temperature," but now this statement is not merely a circumlocution, it is also an error. *V* alone does not tell us the temperature; we must, at the least, also take into account the pressure. To insist that temperature can be represented by volume alone, or by pressure alone, or by any other single thing that can be readily and directly measured, is to mistake the nature of temperature. Just so, I am going to argue that *the simple LR mistakes the nature of evidence*, by obscuring the fact that the evidence itself is not a number, and moreover, that the evidence is not any single thing that can be readily and directly measured, but instead, it is a function of (at least) two measurable things.

(2) The Insidiousness of Simple Hypotheses

To begin with, we need to define *likelihood*:

> "The likelihood, L(H|R), of the hypothesis H given data R, and a specific model, is proportional to P(R|H), the constant of proportionality being arbitrary." Edwards (1992) (p. 9)

Two key points are familiar: (i) likelihood represents a feature of an hypothesis given data, not the other way around; and (ii) likelihood is related to but not the same as probability,

since it is defined only up to an arbitrary multiplicative factor and therefore does not follow the Kolmogorov axioms. I will not rehearse the advantages of likelihood in spelling out a theory of statistical evidence, but suffice it to say that likelihood enables inferences to proceed independently of what are, arguably, extraneous features of study design, including the sampling distribution of all those observations that might have occurred but didn't.

There is a third important feature of this definition as well, and this regards the nature of the *hypotheses* to which the definition is intended to apply. Edwards is, as always, explicit:

> "An essential feature of a statistical hypothesis is that its consequences may be described by an exhaustive set of mutually-exclusive outcomes, to each of which a definite probability is attached." (p. 4)

This precludes consideration of likelihoods involving *composite* hypotheses. For instance, in the context of a coin-tossing experiment in which x independent tosses have landed heads and y have landed tails, and letting θ=P(heads), one can write the likelihood $L(\theta=0.1|x, y)$, or $L(\theta=0.2|x, y)$. These likelihoods involve "simple" hypotheses, in which θ is assigned a single numerical value, so that the corresponding probability $P(x, y|\theta)$ returns a single number on the probability scale for each possible outcome (x, y). But one can *not* write $L(\theta=0.1$ or $\theta=0.2|x, y)$, because the latter involves a "composite" hypothesis, which does not assign a definite probability to the observed outcome. To know the probability of observing (x, y) under the hypothesis "$\theta=0.1$ or $\theta=0.2$," we would need not only to know the probability of (x, y) for each θ, but also, we would need to know the prior probabilities of $\theta=0.1$ and $\theta=0.2$. As these prior probabilities lies outside the likelihood, they are not admissible on the

evidentialist view.

But even the simplest examples of statistical reasoning generally involve hypotheses that appear on the face of things to be composite; e.g., we might be interested in whether the coin is biased toward tails or fair, which would appear to involve the improperly formed hypothesis $\theta<0.5$. This situation is handled by treating composite hypotheses "solely on the merits of their component parts" (Edwards, p. 5). Thus in forming the LR corresponding to 'coin is biased toward tails' vs. 'coin is fair,' we would need to consider separately the (infinitely many) simple LRs in the form $L(\theta=\theta_i|x, y)/L(\theta=0.5|x, y)$, for each possible i^{th} value of $\theta\leq 0.5$. Now the LR is a function of θ, not a single number (Figure 1).

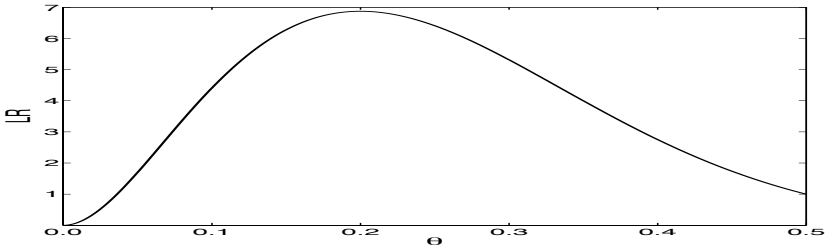


Figure 1 LR as a function of θ for $x = 2$, $y = 8$.

In practice it seems that what is important is not so much the proscription against composite hypotheses, but rather the prescription for how they may be interpreted. We can graph the LR as a function of θ, as if we were admitting composite hypotheses, but we can only make statements like "$\theta=0.2$ is supported over $\theta=0.5$, on given data, by LR=6.9," while

"θ=0.1 is supported over θ=0.5, on those same data, by LR=4.4."[5] But as a practical matter, the graph is not a sufficiently concise summary for general scientific applications. We still need some way to reduce the function LR(θ) to a single number summarizing the strength of the evidence.

And this is where we get into trouble, because focus shifts naturally to the *maximum* LR (MLR), which occurs over the best supported value – the maximum likelihood estimate (m.l.e.) – of θ. Indeed, given that we are only allowed to make statements about one simple hypothesis comparison at a time, the MLR, itself a ratio of two simple likelihoods, appears as the best single constituent LR to use as a summary feature of the LR graph. (Below I consider how relaxing the requirement that hypotheses must be simple frees us up to consider other features.) We have now successfully summarized the *function* LR(θ) as a single number, the MLR, but this summary is tethered to the m.l.e.. We appear to have answered the question: How well supported is the m.l.e. compared to (one or more individual) alternative values of θ? But that is not the question we asked initially, which was about the evidence.[6]

The m.l.e. of θ arrives on the scene as a seemingly innocuous point of special interest, the value that corresponds to the maximum support, but it rapidly takes over, embroiling us in a downward spiral of increasingly perplexing difficulties. One immediate issue with relying on the MLR to summarize the evidence (continuing to focus for ease of discussion on the coin-

[5] Moreover we can only make such statements when both the data and the form of the likelihood are the same in the numerator and the denominator of the LR, for only in such cases will the constants of proportionality cancel.
[6] Hacking (p. 28 ff.) makes clear the conceptual reasons for keeping estimation and evidence (or support) separate.

tossing example, in which maximization occurs only in the numerator of the LR), is that MLR≥1: the MLR can only show evidence in favor of the numerator but never in favor of the denominator. This is problematic, like using a thermometer in which the mercury is prevented from receding.

Another problem with the MLR is that it begs the question of measurement scale in a particularly obvious way, because its evidential meaning would appear to require some kind of adjustment to compensate for the maximization itself. The more parameters we maximize over (again, for ease of discussion, assuming maximization occurs only in the numerator), the larger the MLR becomes. How are we to separate the portion of the MLR reflecting the evidence from the portion representing an artifact of the process of maximization? It becomes particularly hard to retain the fiction that the numerical value of the *maximum* LR has some prima facie meaning with respect to the underlying evidence, regardless of the number of parameters over which the LR is maximized.

There is a third, more subtle but at least as damaging, difficulty with summarizing evidence via MLRs. Simple LRs can be multiplied across two data sets, but MLRs can not be multiplied. Rather, to obtain the MLR based on two sets of data, we first combine the data to find the new m.l.e., which is a kind of weighted average of the two original m.l.e.s, and then we find the new MLR with respect to this average m.l.e. on the combined data. Now consider a situation in which data set D_1 favors H_2 by some substantial amount, and D_2 also favors H_2, but by a lesser amount. In such situations it is not uncommon for the combined support for H_2 to be less than the original support on D_1 alone. But this is not how *evidence* behaves:

strong evidence for H_2 followed by weaker evidence also supporting H_2 ought to lead to *stronger* evidence for H_2, not intermediate evidence. (A blood type match following a DNA match does not lessen the evidence that the defendant was at the crime scene.[7]) This means that we cannot in practice differentiate between situations in which new data are truly diminishing the evidence, and situations in which the evidence is in fact increasing but the MLR at the average m.l.e. goes down anyway. This tendency of the MLR to "average" across combined data is entirely due to its dependence on the m.l.e.; simple LRs do not share this defect.[8]

Of course none of this need surprise unreconstructed evidentialists, who, after all, disavowed composite hypotheses – and therefore any need for maximization – from the start. But then beyond the simplest of examples, we are left with an irreducible graph of the component simple LRs, not a single number. This is true already in single-parameter cases; the problem is only exacerbated in higher dimensions.

There is also the matter of masking the nature of the real problem: by focusing initially only on those situations in which the LR is a single number, we missed Hacking's *measurement* question, how do we ensure that this number always 'means the same'? It is only when we consider composite hypotheses that it becomes clear we were never warranted

[7] This example was suggested by Hasok Chang.
[8] This issue plays a salient role in the current "crisis" of non-replication of statistical findings in the biomedical and social sciences, where the tendency of p-values and MLRs to "regress to the mean" upon attempts to replicate initial findings is widely interpreted as meaning that the evidence has gone down. In the absence of a properly behaved evidence measure, however, this conclusion is entirely unwarranted.

in the first place in assuming that the face value of the LR for a simple vs. simple hypothesis comparison *is* the evidence. Composite hypotheses force us to think in terms of the LR graph, which, precisely because it is not a single number, immediately raises the issue of which *feature(s)* of the graph might be relevant to the evidence. Composite hypotheses are crucial, not only because they are scientifically relevant, but also, because they beg a question all but hidden as long as we focus only on simple hypotheses.

The urge to sidestep the problem of the evidential interpretation of the MLR is the reason evidentialists have been reluctant to admit composite hypotheses into their formalism in the first place. But it is fair to say that they have failed to provide any viable alternative to the MLR as the summary measure of evidence strength in practice. The preoccupation with simple hypotheses has entailed inherent difficulties for the program, and it has also masked a basic underlying calibration issue. The good news, I believe, is that it has also been masking the possibility of a solution.

(3) Towards a Solution to the Measurement Calibration Problem

Consider again the coin-tossing experiment and LR(θ) as shown in Figure 1. Let us suppose, following the spirit if not the letter of the Law of Likelihood, that all of the evidential information is captured, somehow, in this graph. What *feature(s)* of the graph should we take as representing the degree of evidence?

The MLR of course is one possibility, but I have already stated some objections to this option. An alternative would be to use the *area* under the graph (ALR). (Note that this is

only possible if we allow ourselves to consider the truly composite hypothesis $\theta<0.5$, because the ALR requires simultaneous consideration of all of the constituent simple hypotheses.[9]) But while we're at it, why not also consider using *sets of features* of the graph? For instance, the evidence might be a function of the both the MLR and the ALR, e.g., their product, or their ratio. What we need is a methodology for figuring out which among the many possibilities is the correct one.

The methodology I propose is quite simple, at least to begin with. Let's consider the *behavior of candidate evidence measures* in situations where we have clear intuitions regarding the *behavior of evidence*, and see which of our candidate measures behaves like the object of measurement, the evidence. Here I will illustrate using coin-tossing "thought experiments" to discover patterns of behavior of the evidence with changes in data, considering the evidence that the coin is either biased toward tails or fair. I propose that, perhaps with a little persuasion, I could convince you that the following patterns capture *what we mean* when we talk about statistical evidence in this context. (Here I summarize the data in terms of n=the number of tosses, and x/n=the proportion of tosses that land heads.)

(i) Evidence as a function of changes in n for fixed x/n For any given value of x/n, the evidence increases as n increases. The evidence may favor bias (e.g., if $x/n = 0.05$) or no bias (e.g., if $x/n = ½$), but in either case it gets stronger with increasing n.

[9] The ALR is proportional in this simple example to the Bayes factor under a uniform prior on θ, which is sometimes interpreted in Bayesian circles as a measure of evidence strength; it is also proportional to the relative belief (Evans 2015), another Bayesian proposal for measuring evidence. But the ALR itself does not involve a prior, so I see no prima facie reason for the evidentialist to balk at this suggestion, once composite hypotheses are allowed.

(ii) Evidence as a function of changes in *x/n* for fixed *n* If we hold *n* constant but allow *x/n* to increase from 0 up to, say, 0.20, the evidence favoring 'coin is biased' diminishes: i.e., the evidence for bias is stronger the further *x/n* is from ½. But we have also already noted that when *x/n* is close to ½ the evidence favors 'coin is fair.' Therefore, as *x/n* continues to approach ½, at some point the evidence will shift to favoring 'coin is fair,' and from that point, the evidence for 'coin is fair' will increase the closer *x/n* is to ½.

(iii) Rate of evidence change as a function of changes in *n* for fixed *x/n* For given *x/n*, as *n* increases the evidence *increases more slowly* with fixed increments of data. E.g., consider evidence in favor of bias with one additional tail (T), following T, or TT, or TTT. When the number of tails in a row is small (i.e., when there is weak evidence favoring bias), each subsequent T makes us that much more suspicious that the coin is biased. But suppose we have already observed 100 Ts in a row: now one additional T changes our sense of the evidence hardly at all, as we are already quite positive that the coin is not fair.[10]

(iv) *x/n* as a function of changes in *n* (or vice versa) for fixed evidence It follows from (i) and (ii) that in order for the *evidence* to remain constant, *n* and *x/n* must adjust to one another in a compensatory manner. E.g., if *x/n* increases from 0 to 0.05, in order for the evidence to remain the same *n* must increase to compensate; otherwise, the evidence would go down, following (ii) above. By the same token, it is readily verified that if (i)

[10] This underscores the point made above that evidence is not inherent in the data (say, a single toss T), but rather, evidence is a relationship between the data and the hypotheses that depends on context.

and (ii) hold, then as x/n continues to increase, at some point n must begin to decrease in order to hold the evidence constant as the evidence shifts to favoring 'coin is fair.'

Note that at this point we have not mentioned probability distributions, likelihoods, or parameterization of the hypotheses. These patterns characterize evidence in only a very informal, vague manner. However, by the same token, they exhibit a kind of generality: they derive from our general sense of evidence, from what we *mean* by statistical evidence before we attempt a formal mathematical treatment of the concept.

Can we find a precise mathematical expression that exhibits these patterns? As illustrated in Figure 2, the *ratio RLR=MLR/ALR exhibits all of the expected behaviors*. By contrast, neither MLR nor ALR shows all four of these patterns. For instance, MLR, as already noted, cannot show increasing evidence in favor of H_2 because it can never favor H_2 in the first place; and both MLR and ALR increase exponentially in n for fixed x/n rather than showing the concave-down pattern in 2(a).

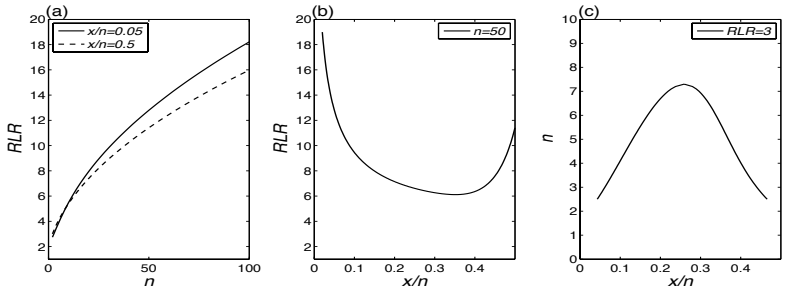


Figure 2 Patterns of behavior of RLR for coin-tossing thought experiments: (a) Patterns (i) and (iii); (b) Pattern (ii); (c) Pattern (iv).

Of course none of this proves that RLR is the correct, or optimal (or properly calibrated) measure of evidence. But this style of reasoning buys us an important methodological tool. Whichever features of the LR graph we consider and however we combine them, we must be able to show that the resulting evidence measure *behaves like the evidence*. When proposing candidate evidence measures anything goes, but only those candidates that behave appropriately remain on the ballot. And even in this very simple example, two obvious candidates – the MLR and the ALR – have already dropped out of contention.

Of course, there is no reason to assume that what works in this simple case (RLR) will work in more complicated cases, nor have we yet resolved the ECP's fundamental calibration issue. Establishing that a measure behaves like the object of measurement is only a first step, but it is a vital step not previously taken. It provides an "empirical" measurement scale, not an absolute scale, much as early thermoscopes provided good experimental tools while falling short of proper, absolute, calibration (Chang 2004).[11] Projecting an empirical measure onto an absolute scale requires a broader theoretical foundation, but one needs the empirical measure first. My point here is simply that confronting the ECP head on, and in the context of composite hypotheses, opens the door for the first time to the possibility of establishing a proper measurement scale for statistical evidence.

Note too that the coin-tossing exercise suggests the existence of an *equation of state* involving the three quantities (n, x/n and the evidence), such that fixing any one quantity

[11] Indeed, the ECP poses what Chang calls a "nomic" measurement problem, much like the nomic problem of temperature measurement. What I am describing here is a necessary but not sufficient stage in resolving a nomic problem.

while allowing a second one to change requires a specific compensatory change in the third. This in turn suggests a new, and potentially very powerful, way to think about the laws governing the behavior of LRs. I'm not aware of any evidentialist work that considers such equations, but I see no reason that an evidentialist-at-heart should be prohibited from pursuing their study.

(4) Relaxing the Foundations To Include Composite Hypotheses

In order to tackle the ECP in the terms of the preceding section, we need to amend the foundations of evidentialism, but only slightly. I propose the following changes. First, let's retain Edwards definition of likelihood, as quoted above, but insert the word "simple" (which is tacit in Edwards' original statement): "The likelihood, L(H|R), of a *simple* hypothesis H given data R, and a specific model, is proportional to P(R|H), the constant of proportionality being arbitrary." Second, we can again add the word "simple" to his characterization of a statistical hypothesis: "An essential feature of a *simple* statistical hypothesis is that its consequences may be described by an exhaustive set of mutually-exclusive outcomes, to each of which a definite probability is attached." But we can now add a definition of likelihood for a composite hypothesis: "A *composite* hypothesis H given data R, and a specific model, is the set of all constituent simple hypotheses, defined up to a single constant of proportionality." Thus the essential feature of a *composite* hypothesis is that *each of its constituent simple hypotheses* may be described by an exhaustive set of mutually-exclusive outcomes, to each of which a definite probability is attached. We can now use this definition

of a composite hypothesis to define the corresponding composite likelihood, as the set of all constituent simple likelihoods.

Under my proposal, the spirit of the Law of Likelihood can be retained: We can say that all of the *evidential information* conveyed by given data regarding a comparison between two hypotheses on a particular model is contained in the LR, where, under the expanded definition of hypotheses, the LR is understood to be a function of all unknown parameters, or better still perhaps, a *graph*. This can equivalently be read as a definition of *evidential information,* as whatever changes the LR graph.[12] But the idea that the (simple) LR itself expresses the degree or weight of the evidence must be abandoned. What I have attempted to argue here is that there is at least the possibility of replacing this notion with something more useful.

Discussion

Evidence is a general and vague term in science. Statistical evidence is a narrower concept, but it still inherits some of this vagueness. One way to tackle a general and vague term is by seeking a precise definition that maintains full generality, but of course, this might not be possible. Weyl (1952) has suggested another approach:

> "To a certain degree this scheme is typical for all theoretic knowledge: We begin with some general but vague principle, then find an important case where we can give that

[12] I borrow this idea from Frank (2014), who defines *information* as whatever changes a probability distribution.

> notion a concrete precise meaning, and from that case we gradually rise again to generality… and if we are lucky we end up with an idea no less universal than the one from which we started. Gone may be much of its emotional appeal, but it has the same or even greater unifying power in the realm of thought and is exact instead of vague." (p. 6)

Can evidentialism be redeemed and made truly useful to science? Of course I have not proved that the answer is yes. But in section (3) I illustrated a case in which we appear to be able to give the vague concept of statistical evidence a concrete, precise meaning, via the quantity RLR=MLR/ALR. It remains to be seen whether it is possible to rise again to generality from this first step. But for those of us who agree with most of what Barnard, Hacking and Edwards have to say on the subject, it seems worthwhile to see how far we can take this line of reasoning. This also seems to be a singular opportunity for philosophers of science to step into the breach and at least *try* to solve a problem that has long stood between one of the needs of science – for well-behaved quantitative measures of evidence – and the capabilities of conventional statistical methodologies.

References

Barnard G.A. "Statistical Inference." *J Royal Stat Soc* XI, no. 2 (1949):115-39.

Chang H. *Inventing Temperature: Measurement and Scientific Progress*. New York:Oxford UP, 2004.

Edwards A.W.F. *Likelihood*. Baltimore:Johns Hopkins UP, 1992. Orig. Cambridge UP, 1972.

Evans M. *Measuring Statistical Evidence Using Relative Belief*, Monographs on Statistics and Applied Probability. Boca Raton:CRC Press, Taylor & Francis Group, 2015.

Forster M, Sober E. "Why Likelihood?" In *The Nature of Scientific Evidence*, Taper & Lele eds., 153-90. Chicago:Chicago UP, 2004.

Frank S.A. "How to Read Probability Distributions as Statements About Process." *Entropy* 16(2014):6059-98.

Good I. J. *Probability and Weighing of Evidence*. London:Griffon, 1950.

Hacking I. *Logic of Statistical Inference*. London:Cambridge UP, 1965.
———. "Review of Edwards' Likelihood." *British J Phil of Sci* 23(1972): 132-37.
Royall R. *Statistical Evidence: A Likelihood Paradigm*. London:Chapman & Hall, 1997.
Weyl, Hermann. *Symmetry*. Princeton UP, 1952.

What Basic Emotions Really Are

Encapsulated or Integrated?

Abstract: While there is ongoing debate about the existence of basic emotions (BEs) and about their status as natural kinds, these debates usually carry on under the assumption that BEs are encapsulated from cognition and that this is one of the criteria that separates the products of evolution from the products of culture and experience. I aim to show that this assumption is entirely unwarranted, that there is empirical evidence against it, and that evolutionary theory itself should not lead us to expect that cognitive encapsulation marks the distinction between basic and higher cognitive emotions. Finally, I draw out the implications of these claims for debates about the existence of basic emotions in humans.

1. Introduction

It is widely held among emotion theorists that there is some theoretically interesting distinction between basic and higher cognitive emotions. On this picture, basic emotions (BEs) are primarily structured by evolution whereas higher cognitive emotions are substantially structured by either culture or individual experience. While there is ongoing debate about the existence of BEs and about their status as natural kinds, these debates usually carry on under the assumption that BEs are encapsulated from cognition and that encapsulation is one of the criteria that separates the products of evolution from the products of culture and experience. I aim to show that this assumptions is entirely unwarranted, that there is empirical evidence against it, and that evolutionary theory itself should not lead us to

expect that cognitive encapsulation marks the distinction between basic and higher cognitive emotions. Finally, I draw out the implications of these claims for the existence of basic emotions in humans.

In the following section, I characterize the received view of BEs, which holds (among other things) that BEs are solutions to *basic life problems* in our evolutionary past. Then I consider and reject some of the reasons to think that BEs are cognitively encapsulated. In the second section, I provide an example of a BE in rodents that bears the marks of cognitive integration (as opposed to encapsulation). The basic life problem that likely shaped this emotion appears to demand substantial cognitive integration. In the third section, I draw out the implications for a current debate in emotion theory concerning the existence of BEs in humans.

2. Basic Emotions

BEs – including anger, fear, happiness, sadness, disgust, and surprise (for an extended list, see Ekman & Cordaro, 2011) – are thought to be human-typical behavioral syndromes that include involuntary facial expressions of emotion, physiological changes (e.g. in heart rate, blood pressure, and hormone levels), and changes in bodily posture (including bodily social displays and orienting responses). According to BE theory, these syndromes have a similar kind of evolutionary explanation and similar neural and psychological mechanisms. Specifically, they each evolved to address basic life problems or adaptive problems (such as

resource competition, avoidance of predators and avoidance of poisons and parasites). Some of these basic life problems are ones that we share with non-human animals.

Moreover, the elicitation and production of these syndromes (including the coordination of various response components) are supposed to be explained by *automatic appraisal mechanisms* and *affect programs*, respectively (Ekman, 1977, 1999). For instance, affect programs explain phenomena observed in experiments that ask people to distinguish photographs of facial expressions of emotions, connect these expressions with emotion terms, or rate their appropriateness in response to vignettes (for an overview, see Ekman, 2003). They are also supposed to explain the results of experiments that connect facial expressions with changes in physiological response components (Ekman, Levenson, & Friesen, 1983; Levenson, Ekman, & Friesen, 1990). To generalize, affect programs are introduced to explain the observed coordination of various response components and the cross-cultural production of these various syndromes (which is thought to explain widespread recognition of facial expressions across cultures).

3. Unwarranted Assumptions Concerning Cognitive Integration

Many emotion theorists claim that BEs lack cognitive integration. In this section, I argue that these claims are based on unwarranted assumptions.

Assumption 1: Cognitively Integrated only if Informationally Integrated

In most cases, questions about the integration of emotions with cognition concern the possibility that emotions are modular in Fodor's (1983) sense. This depends (among other

things) on whether they can store *information* that cognitive systems cannot access (*informational encapsulation*); or whether *information* from other cognitive systems can interfere with the operations of an emotion (*cognitive penetrability*); or whether people have conscious access to emotional processes or merely their outputs (*opacity*); or whether the *information* that an emotion provides is general as opposed to specific (which would imply *shallow outputs*). These are some of the more well-known marks of cognitive integration or its absence, encapsulation.

Philosophers and psychologists alike usually proceed under the assumption that integration with cognition depends entirely on whether information is integrated in these ways. These assumptions translate to discussions about BEs, where evidence for lack of *informational* integration is sometimes used as evidence for lack of *cognitive* integration *simpliciter*:

> Three other types of evidence suggest that [basic] emotion processes can operate independently of cognition. Emotions have been induced by unanticipated pain…, manipulation of facial expressions…, and changing the temperature of cerebral blood… In all these conditions the immediate cause of the emotion was noncognitive. (Izard, 1992, p. 563, see also his 2007)

Here, Izard apparently assumes that the impenetrability of BEs constitutes evidence that BEs operate independently of cognition. The fact that they respond to low level inputs or processes to which other systems have limited access certainly suggests that emotional states can respond to information that is not integrated with cognition. In addition, there is evidence

that people cannot fully control facial expressions of BEs (Ekman, 1972; Friesen, 1973), suggesting that BEs are cognitively impenetrable. Overall, BEs appear to lack informational integration.

Nevertheless, the realm of the cognitive picks out not only informational states, but also includes a broader range of internal states that function as causal intermediates between stimulus and response, perception and action (Rey, 1997). Cognitive states so understood include not only informational states (such as beliefs) but also motivational states (such as desires). Moreover, questions about cognitive integration may be asked about either informational or motivational states. If so, the possibility arises that the two forms of cognitive integration are independent of one another. If so, any inference from the one to the other is invalid.

This becomes clear when we consider hunger. Hunger may very well be akin to desire (a paradigmatic case of a cognitively integrated state) in the sense that it can interact with other cognitive systems to produce flexible or novel behaviors, as when rodents take novel "short cuts" to get to a food box in a maze (Olton, 1979; Tolman, 1948). Short cut behaviors suggest that hunger is a motivational state that can incline rodents to the pursuit of an end (e.g. food consumption) by selecting from a range of different means, perhaps by interacting with informational states that relate means to ends (e.g. means-ends beliefs). Even so, hunger may be cognitively impenetrable in that it may be triggered by low level stimuli and processes (e.g. low-level detection of changes in blood sugar). Moreover, when one feels hungry, one cannot interfere with the feeling of hunger by thinking about it (e.g. by noticing

that the amount of energy one's body has stored in fat deposits is more than enough to sustain oneself). One can even imagine that it is informationally encapsulated: it might store information (e.g. about which foods are more calorically dense) that other systems cannot directly access.

These conceptual possibilities suggest that questions concerning the integration of informational states are conceptually independent of questions concerning the integration of motivational states. Hunger may be informationally encapsulated while retaining a degree of integration as a motivational state. Wholesale encapsulation, therefore, does not follow from informational encapsulation. If this is correct, then inferences like the one Izard draws above are invalid: having non-cognitive inputs is not a reason to think that emotions operate independently of cognition. They might very well operate in concert with cognition on the output side or as motivational states. Before I raise that possibility, consider another reason to rule it out at the outset: that BEs are not integrated with propositional attitudes, including beliefs *and* desires.

Assumption 2: Integration with Beliefs and Desires is the Criterion for Cognitive Integration

Contrary to the previous assumption, this one respects the distinction between motivational and informational integration. Nevertheless, I argue that it sets the bar for cognitive integration too high.

Isaac Wiegman
10/19/2016

To see this, consider Griffiths' (Griffiths, 1997, 2004) views on the distinction between basic and higher cognitive emotions. First, he draws on some of the same evidence as Izard to conclude that BEs are opaque and informationally encapsulated. Since they have these and other marks of modularity, Griffiths thinks BEs have "limited involvement" with higher cognitive processes, which are "…the processes in which people use the information of the sort they verbally assent to (traditional beliefs) and the goals they can be brought to recognize (traditional desires) to guide relatively long-term action and to solve theoretical problems." (Griffiths, 1997, p. 92) Here, Griffiths may be making the same faulty assumption as Izard (that informational encapsulation implies cognitive encapsulation more broadly). However, let us grant that he may have additional reasons to think that emotions are not integrated on the output side or qua motivational states.

From this, Griffiths draws a broader conclusion: that BEs are not "flexible [or] integrated with long-term, planned action" and are instead "restricted to short-term, stereotyped responses" (Griffiths, 1997, p. 241). The apparent assumption is that if BEs are not integrated with beliefs, desires and long-term planning, then the only alternative is that they are similar to fixed action patterns, being inflexible and stereotyped. Griffiths makes no explicit argument for this assumption, perhaps at the time it was widespread enough to make further argument otiose.

Nevertheless, it has become a tendentious assumption for several reasons. First, the phenomena of intelligent action are much broader than deliberate, "long-term, planned action" mediated by beliefs and desires. For instance, Ginet (1990) argues that many clear

cases of actions (as distinct from mere behaviors, such as reflexes or fixed action patterns) are not plausibly mediated by conscious beliefs, desires or intentions: involuntarily crossing one's legs, kicking a door in anger, impulsively pulling a loose thread from one's clothes, and slamming on the brakes to avoid hitting a dog. These actions are not mere behaviors or reflexes. That is, they appear to be purposive and guided by the agent, but it is difficult to find belief-desire style explanations that render them intelligible.[1] Why not think that BEs can influence actions more akin to this variety than to "long-term, planned actions"? Griffiths never raises this question, neither does he give reason to rule out the possibility that BEs cause actions intermediate between long-term planned action and stereotyped behavioral responses.

Second, if we ask what might explain the other varieties of action that Ginet picks out, it may be that such actions are guided by other representational states, aside from conscious or verbally reportable beliefs, desires and intentions. For instance, in the last twenty years, cognitive scientists have begun to emphasize the role of unconscious or non-conceptual representational states in generating flexible and intelligent behavior (Bermúdez, 2003). Informational states aside from beliefs include perceptual representations, map-like spatial representations and representations of affordances. Motivational states aside from desires include drives, incentives and feedback mechanisms.

[1] See also Hursthouse (1991).

Isaac Wiegman
10/19/2016

The flexibility and intelligence of these representational states becomes clear when we consider animal behavior. Nonhuman animals display forms of intelligent or purposive or instrumental behavior (see e.g. Balleine & Dickinson, 1998), even while lacking linguistically mediated propositional attitudes. This suggests that instrumental behaviors in non-human animals are underwritten by a different form of cognitive integration. Consider what Susan Hurley calls *holistic flexibility*:

> The holistic flexibility of intentional agency contributes a degree of generality to the agent's skills: a given means can be transferred to a novel end, or a novel means adopted toward a given end. The end or goal functions as an intervening variable that organizes varying inputs and outputs and allows a degree of transfer across contexts. (Hurley, 2003, pp. 237–38)

Where this sort of flexibility is found, it suggests that behavior is best explained with reference to informational states which represent the means available to an organism (e.g. affordances) and motivational states that represent its ends (e.g. drive states), which can interact interchangeably in order to bring about the same end by various means or to deploy a single means to bring about various ends.

Nevertheless, these informational and motivational states may sometimes lack inferential integration with beliefs and desires. Even in humans, phenomena like "blind-sight" suggest that perceptual representations can flexibly guide behavior without being integrated with verbally reportable states. That is, even though these perceptual states are not verbally reportable or consciously accessible, these informational states mediate goal-

directed behaviors (e.g. putting a plate in a slot) rather than just reflexes and fixed action patterns (see e.g. Goodale, Milner, Jakobson, & Carey, 1991). All this suggests that Griffiths' requirements on cognitive integration are too stringent. Verbal reportability and conscious accessibility of a representational state is not necessary for such a state to influence flexible behaviors. To my knowledge there is no evidence that BEs fail to meet less stringent requirements on cognitive integration such as holistic integration.

Once the full range of representational states is expanded in this way (beyond beliefs and desires), it becomes possible that BEs have some degree of motivational integration with other representational states aside from conscious beliefs and desires to produce behaviors that are more flexible and purposive than stereotyped behaviors. Griffiths provides no reason to rule out this possibility.

4. Evidence of Integration in a Basic Emotion

In fact, there is some reason to rule it in. Consider the instinctive patterns of territorial behavior of rodents. These behaviors have been investigated in great detail using a resident-intruder experimental paradigm (for an overview, see D. C. Blanchard & Blanchard, 1984, 2003) add it Adams RRR) in which resident (who have occupied a cage or colony for a few weeks) will attack unfamiliar male intruders introduced into their cage. The attacks of the resident and the defensive maneuvers of the intruder comprise sets of stereotyped behaviors. Each attack behavior of the resident is paired with a matching defensive maneuver of the intruder. The resident adopts a set of stereotyped postures and attacks aimed at biting the

Isaac Wiegman
10/19/2016

dorsal surfaces of the intruder. On the other hand, the intruder adopts a distinctive set of stereotyped behaviors aimed at avoiding or blocking the resident's attempts to bite its back.

While these behaviors are certainly stereotyped, they are not brittle or reflexive. For instance, attacks of residents vary depending on the defensive strategy adopted by the intruder, and they seem to be governed by a motive to approach and attack that persists the entire time that the intruder is present. By contrast, the intruder rat's whole suite of behaviors seems to be governed by a persistent motive to escape and avoid.

Isaac Wiegman
10/19/2016

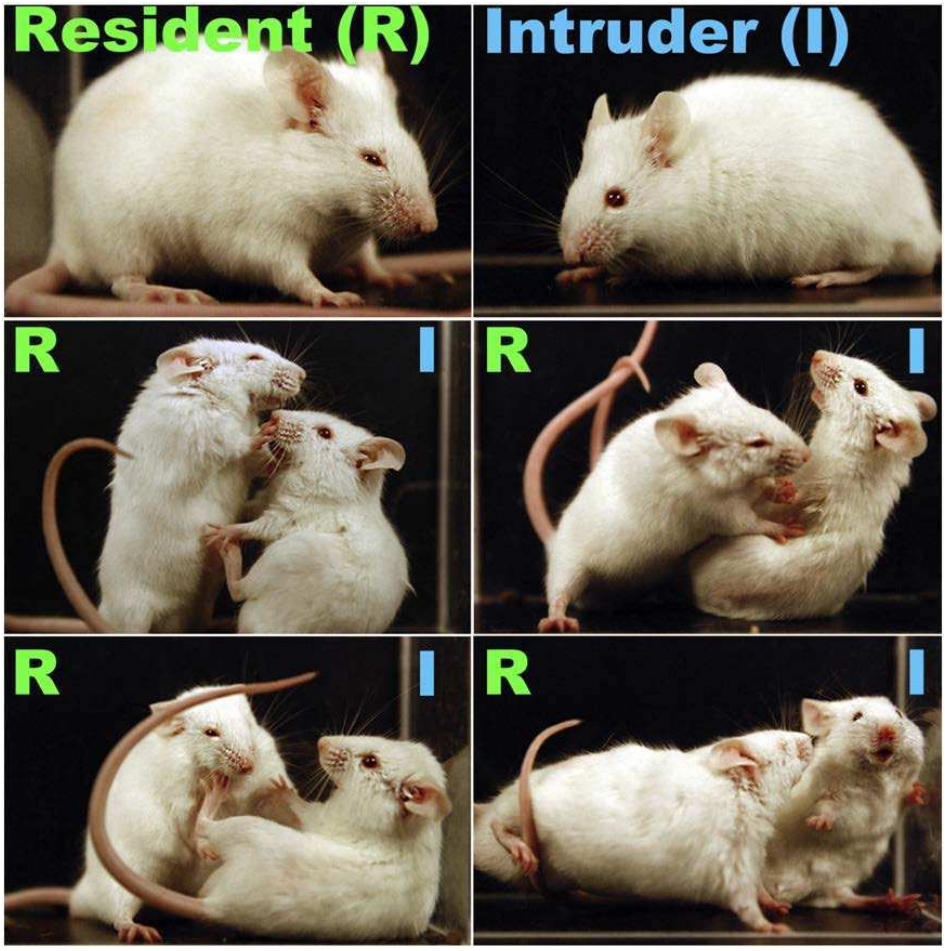


Figure 1 Confrontation and avoidance behaviors (e.g. facial expressions, postures and maneuvers) of resident and intruder mice (respectively). From Defensor and Corley (2012), p. 683 permission pending © Elsevier. Originally published in Physiology and Behavior.

Isaac Wiegman
10/19/2016

What scientists have discovered about these behaviors (the flexibility of these behaviors and their coherent aims) indicates that they are produced by two underlying motivational systems, what I call the confrontation and avoidance systems (D. C. Blanchard & Blanchard, 1984, 2003; D. C. Blanchard, Litvin, Pentkowski, & Blanchard, 2009). The confrontation system is tuned to bring about a specific end state, repeated back-biting. Moreover, this motive does not depend on learning: rats which have been socially isolated from birth will still attempt to bite the back of an intruder (Eibl-Eibesfeldt, 1961). So far, the focus has been on cases in which a given rodent is purely motivated by confrontation or avoidance, but aggressive encounters in the wild usually involve a mix of offensive and defensive postures. This suggests that these motivational systems can be activated simultaneously or in close succession to produce mixed patterns of behavior.

Regardless, these systems have many of the characteristics of affect programs in humans. They are posited to explain a coordinated suite of behaviors and physiological changes that may include facial expressions, cardiovascular changes, and endocrine responses (Defensor, Corley, Blanchard, & Blanchard, 2012; Fokkema, Koolhaas, & van der Gugten, 1995). Moreover, these systems are tailored to solve basic life problems. Specifically, the confrontation system solves the problem of defending territories from other males for breeding purposes (and without fatally injuring kin in the process), whereas the avoidance system solves the problem of avoiding occupied territories and failing that, defending against the attacks of residents. For these reasons, we have all the same reasons to

postulate BEs in rodent that we have in humans. Let us suppose then that the confrontation and avoidance systems are BEs in rodents.

Interesting for my purposes, under certain conditions, the presence of the unfamiliar male can produce highly flexible and novel behaviors. In the bound-intruder task, an intruder is tied down on a Plexiglas plate with only its ventral surfaces (belly-side) exposed and placed in the cage of a resident, so that the resident cannot easily bite the back of the intruder. As a result, the resident will sometimes bite at the bands that tie down the intruder or dig under the intruder so that the resident can bite the intruder's back (R. J. Blanchard, Blanchard, Takahashi, & Kelley, 1977). In contrast, none of these behaviors are adopted when the intruder is tied down with his back exposed.

These instrumental behaviors are clearly not stereotyped forms of attack, rather they are forms of flexible behavior adjustment to achieve the aim of biting the intruder's back: they exhibit holistic integration. In this case, the same end can be achieved by several, novel means. Attempts to bite the intruder's bonds or to dig underneath the intruder are novel means toward the end of biting the back of the intruder. Moreover, some of a resident's means can be deployed toward novel ends. Digging is an element of the rat's behavioral repertoire that is ordinarily used for an entirely different purpose: constructing burrow systems for shelter and nesting (Boice, 1977). This suggests that there are informational states, representations of means (e.g. motor representations of digging, biting, lateral attack, etc.), that can interact interchangeably with motivational states, representations of various ends (e.g. nesting, back-biting, eating etc.), in order to produce flexible behaviors.

Isaac Wiegman
10/19/2016

Importantly, the confrontation system seems to be involved in coordinating flexible back-biting behavior. Moreover, this is something we would predict if it is a solution to the basic life problem of defending a territory from intruders. Flexibility is required to successfully repel an intruder because it is not in the intruder's best interest to be repelled easily or to act predictably. For instance, the intruder would be sure to fare poorly if it acted in a way that accommodates the attacks of the resident. So a single fixed action pattern or even a whole suite of fixed action patterns on the part of the resident would not tend to be successful against the most likely strategy of the intruder. It is more adaptive to have a flexible motivational state that leads to repeated back biting across a wide range of strategies or postures that the intruder might adopt. Rather than leading only to inflexible, stereotyped responses, it appears that solutions to basic life problems sometimes require some degree of motivational integration.

5. Implications for Emotion Theory

If we understand BEs in this way, this changes the shape of an ongoing debate in emotion theory concerning the existence of BEs in humans. In the past, this debate has carried on under the assumption that if an emotion is biologically basic, then one should predict that the various response components of the emotion will have a high degree of coherence; that for example "all instances of anger should have a characteristic facial display, cardiovascular pattern, and voluntary action that are coordinated in time and correlated in intensity."

Isaac Wiegman
10/19/2016

(Barrett, 2006, p. 29) This high degree of coherence is not observed across many emotions (Gentsch, Grandjean, & Scherer, 2013; Reisenzein, Studtmann, & Horstmann, 2013). For instance, when anger is elicited in experimental settings, it is uncommon to observe facial expressions in conjunction with the other putative components of BE anger.

One way of defending the basicality of an emotion against this criticism is to reassess what patterns of emotional response are predicted by BE theory. As we saw in the section above the motivational component of a basic emotion can select novel, instrumental behaviors. Moreover, the motivational component can be indispensable for solving a basic life problem. I think we can add to this the possibility that other response components are not as indispensable as the motivational state. To see this, suppose that anger in humans is a solution to basic life problems of deterring conspecifics from challenges and insults. If so, it may be that the only reliable requirement of successful deterrence (at least in our lineage) is a flexible motivation to retaliate against perceived wrongs (e.g. McCullough, Kurzban, & Tabak, 2012). For instance, a reliable disposition to garner a reputation for revenge (e.g. by avenging personal offenses) appears to be a highly reliable strategy for deterrence (e.g. Daly & Wilson, 1988; Frank, 1988), perhaps more so than any facial expression or physiological responses. If revenge can be served cold, then anger may not always require anything more than a motivation to avenge. If so, then we might *expect* that the only reliably occurring component of anger is the relevant motivational state. But if this is correct, then evidence of low coherence is not evidence against the existence of BE anger. While this is a just-so story that may or may not end up being true, it shows that the expected level of coherence in a BE

depends on which basic life problem shaped that emotion. In some cases, we might expect the motivational state to be the only component that does not significantly vary across the situations in which these problems arise. In that case, contextually variable responses will be the norm rather than the exception.

6. Conclusion: What Basic Emotions Really Are

So what are basic emotions? Like other theoretical terms, part of the theoretical function of basic emotions is to place selective stress on competing theories (e.g. Kroon, 1985). In this case, BEs and competing conceptions of emotion allow us to discriminate between evolutionary theories of emotion in competition with radical social constructivist theories (e.g. Barrett, 2014; Lindquist, Siegel, Quigley, & Barrett, 2013).

BEs help distinguish these theories by specifying an architecture for emotion production predicted by evolutionary considerations. The distinguishing factor is whether emotion production is categorical or dimensional (see figure 2). If each BE is a solution to a different basic life problem, then when a BE is elicited, we should see emotional responses that are relevant to that basic life problem and distinct from the responses manifested by other BEs. Emotion production is categorical in the sense that the behavioral responses are controlled by a single emotional state (as distinct from other emotional states that might control a distinct pattern of response). By contrast, if all emotions are socially constructed as

some theorists claim, we might expect to see emotional behaviors controlled directly by multiple dimensions of appraisal (as in the bottom half of figure 2).

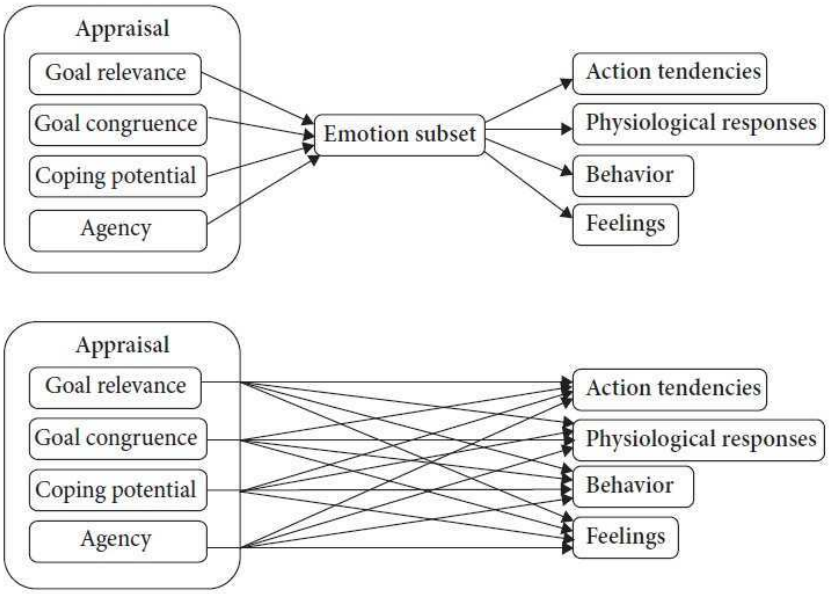


Figure 2 Competing architectures for emotion production. Top diagram is a categorical architecture, whereas the bottom is dimensional. From Moors (2012), p. 266 permission pending © John Benjamins Publishing Company. Originally published in Zachar and Ellis (2012).

Isaac Wiegman
10/19/2016

Until the present, contextual variability of emotional responses has played a decisive role in distinguishing between these two architectures for emotion production. If flexible motivational states are not included among the components of BEs, then discrete emotion production predicts insensitivity to context subsequent to elicitation (though emotion regulation processes can perhaps inhibit or augment emotional responses according to context). However, once flexible motivational states are possible, categorical emotion production is compatible with a greater amount of contextual variability.

Admittedly, this added complexity makes it more difficult to test whether humans have BEs. Nevertheless, it is not impossible. For instance, in the case of anger, researchers have developed a neurological measure of approach motivation (for a review, see Carver & Harmon-jones, 2009). If this motivational state is a component of anger, we can measure whether approach motivation itself is better predicted by contextual variables subsequent to anger elicitation or rather by contextual variables prior to or during elicitation. If contextual variables prior to elicitation do not independently predict approach motivation as BE theory might lead us to expect, then we would have evidence against the existence of BE anger.

I have argued against prevailing assumptions that BEs lack cognitive integration. In the past, evidence against cognitive integration has been concerned with informational integration, and motivational integration has not been considered. Moreover, the assumed requirements for integration concern interaction with verbally reportable or consciously accessible states, and integration with other representational states is ignored. Moreover, BEs in rodents exhibit a form of motivational integration that plausibly hinges on interaction with

Isaac Wiegman
10/19/2016

a wider variety of representational states. Properly understood, BEs are more likely to refer to emotional states in humans.

Isaac Wiegman
10/19/2016

References

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*, 407–419.

Barrett, L. F. (2006). Are Emotions Natural Kinds? *Perspectives on Psychological Science*, *1*(1), 28–58. http://doi.org/10.1111/j.1745-6916.2006.00003.x

Barrett, L. F. (2014). The Conceptual Act Theory: A Précis. *Emotion Review*, 1–20. http://doi.org/10.1177/1754073914534479

Bermúdez, J. (2003). *Thinking without words*.

Blanchard, D. C., & Blanchard, R. J. (1984). Affect and aggression: An animal model applied to human behavior. In R. J. Blanchard & D. C. Blanchard (Eds.), *Advances in the Study of Aggression* (Vol. 1, pp. 1–62).

Blanchard, D. C., & Blanchard, R. J. (2003). What can animal aggression research tell us about human aggression? *Hormones and Behavior*, *44*(3), 171–177. http://doi.org/10.1016/S0018-506X(03)00133-8

Blanchard, D. C., Litvin, Y., Pentkowski, N. S., & Blanchard, R. J. (2009). Defense and Aggression. In G. G. Berntson & J. T. Cacioppo (Eds.), *Handbook of Neuroscience for the Behavioral Sciences* (pp. 958–974). Hoboken: Wiley.

Blanchard, R. J., Blanchard, D. C., Takahashi, T., & Kelley, M. J. (1977). Attack and defensive behaviour in the albino rat. *Animal Behaviour*, *25*, 622–634.

Isaac Wiegman
10/19/2016

Boice, R. (1977). Burrows of wild and albino rats: effects of domestication, outdoor raising, age, experience, and maternal state. *Journal of Comparative and Physiological Psychology*, *91*(3), 649–61.

Carver, C. S., & Harmon-jones, E. (2009). Anger Is an Approach-Related Affect : Evidence and Implications. *Psychological Bulletin*, *135*(2), 183–204. http://doi.org/10.1037/a0013965

Daly, M., & Wilson, M. (1988). *Homicide*. Transaction Publishers.

Defensor, E. B., Corley, M. J., Blanchard, R. J., & Blanchard, D. C. (2012). Facial expressions of mice in aggressive and fearful contexts. *Physiology & Behavior*, *107*(5), 680–5. http://doi.org/10.1016/j.physbeh.2012.03.024

Eibl-Eibesfeldt, I. (1961). The Fighting Behavior of Animals. *Scientific American*, *205*, 112–122.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*. University of Nebraska Press Lincoln. http://doi.org/10.1037/0022-3514.53.4.712

Ekman, P. (1977). Biological and cultural contributions to body and facial movement. In J. Blacking (Ed.), *Anthropology of the body* (pp. 34–84).

Ekman, P. (1999). Basic Emotions. In T. Dalgleish & M. Power (Eds.), *The Handbook of Cognition and Emotion* (pp. 45–60). Sussex: John Wiley & Sons.

Isaac Wiegman
10/19/2016

Ekman, P. (2003). *Emotion Revealed: Understanding Faces and Feelings*. Phoenix Press.

Ekman, P., & Cordaro, D. (2011). What is Meant by Calling Emotions Basic. *Emotion Review*, *3*(4), 364–370. http://doi.org/10.1177/1754073911410740

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, *221*(4616), 1208–1210.

Fodor, J. (1983). *The modularity of mind: An essay on faculty psychology*.

Fokkema, D. S., Koolhaas, J. M., & van der Gugten, J. (1995). Individual characteristics of behavior, blood pressure, and adrenal hormones in colony rats. *Physiology & Behavior*, *57*(5), 857–62.

Frank, R. H. (1988). *Passions Within Reason: The Strategic Role of the Emotions*. New York: Norton. http://doi.org/10.2307/2072516

Friesen, W. (1973). *Cultural differences in facial expressions in a social situation: An experimental test on the concept of display rules*. University of California, San Francisco.

Gentsch, K., Grandjean, D., & Scherer, K. R. (2013). Coherence explored between emotion components: Evidence from event-related potentials and facial electromyography. *Biological Psychology*. http://doi.org/10.1016/j.biopsycho.2013.11.007

Ginet, C. (1990). *On action.*

Goodale, M., Milner, A., Jakobson, L., & Carey, D. (1991). A neurological dissociation

between perceiving objects and grasping them. *Nature*.

Griffiths, P. E. (1997). *What emotions really are: The problem of psychological categories* (Vol. 1997). University of Chicago Press.

Griffiths, P. E. (2004). Emotions as Natural and Normative Kinds, *71*(December), 901–911.

Hurley, S. (2003). Animal action in the space of reasons. *Mind & Language*, *18*(3), 231–256.

Hursthouse, R. (1991). Arational actions. *The Journal of Philosophy*.

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations.

Izard, C. E. (2007). Basic Emotions, Natural Kinds, Emotion Schemas, and a New Paradigm. *Perspectives on Psychological Science*, *2*(3), 260–280. http://doi.org/10.1111/j.1745-6916.2007.00044.x

Kroon, F. (1985). Theoretical terms and the causal view of reference. *Australasian Journal of Philosophy*, (February 2014), 37–41.

Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, *27*(4), 363–384.

Lindquist, K. A., Siegel, E. H., Quigley, K. S., & Barrett, L. F. (2013). The Hundred-Year Emotion War : Are Emotions Natural Kinds or Psychological Constructions ? Comment on Lench , *139*(1), 255–263. http://doi.org/10.1037/a0029038

Isaac Wiegman
10/19/2016

McCullough, M. E., Kurzban, R., & Tabak, B. a. (2012). Cognitive systems for revenge and forgiveness. *The Behavioral and Brain Sciences*, 1–15. http://doi.org/10.1017/S0140525X11002160

Moors, A. (2012). Comparison of affect program theories, appraisal theories, and psychological construction theories. *Categorical versus Dimensional Models of Affect. A Seminar on the Theories of Panksepp and Russell*, 257–278.

Olton, D. (1979). Mazes, maps, and memory. *American Psychologist*.

Reisenzein, R., Studtmann, M., & Horstmann, G. (2013). Coherence between Emotion and Facial Expression: Evidence from Laboratory Experiments. *Emotion Review*, *5*, 16–23. http://doi.org/10.1177/1754073912457228

Rey, G. (1997). Contemporary philosophy of mind: A contentiously classical approach.

Tolman, E. (1948). Cognitive maps in rats and men. *Psychological Review*.

Zachar, P., & Ellis, R. (2012). *Categorical versus dimensional models of affect: a seminar on the theories of Panksepp and Russell*.

Multiple realization and the commensurability of taxonomies

Abstract

The past two decades have witnessed a revival of interest in multiple realization and multiply realized kinds. Bechtel and Mundale's (1999) illuminating discussion of the subject must no doubt be credited with having generated much of this renewed interest. Among other virtues, their paper expresses what seems to be an important insight about multiple realization: that unless we keep a consistent grain across realized and realizing kinds, claims alleging the multiple realization of psychological kinds are vulnerable to refutation. In this paper I argue that, intuitions notwithstanding, the terms in which their recommendation has been put make it impossible to follow, while also misleadingly insinuating that meeting their desideratum virtually guarantees mind-brain identity. Instead of a matching of grains, what multiple realization really requires is a principled method for adjudicating upon differences between tokens. Shapiro's (2000) work on multiple realization can be understood as an attempt to adumbrate such a method.

Multiple realization, *neuroscience*, *autonomy of psychology*, *intertheoretic reduction*

1. Introduction

The multiple realization ("MR") hypothesis asserts, at its baldest, that the same psychological state may be realized in neurologically distinct substrates (Polger 2009). Hilary Putnam's (1967) ingenious suggestion that pain is likely to be a multiply realized kind ("MR kind") rather neatly captures the thought here—while both mammals and molluscs presumably experience pain, they're likely to instantiate it in neurological systems of a very different sort.

MR was played against a popular philosophical theory of mind in the 1960s which attempted to identify mental states with neural states. Since MR implies a many-to-one mapping from neural states to mental states, if it is in fact true that mental states are multiply realized, it follows that no clear identity relation can hold between them. As Bechtel and Mundale (1999, 176) frame the issue, "[o]ne corollary of this rejection of the identity thesis is the contention that information

about the brain is of little or no relevance to understanding psychological processes." When the MR hypothesis first came to prominence, its critics by and large accepted it as empirically correct, and merely denied its touted antireductionist implications. In recent years the debate has struck a new note, with many philosophers calling the empirical hypothesis itself into question. Bechtel and Mundale's (1999) influential paper, followed quickly at the heels by Shapiro's (2000) penetrating analysis of functions, perhaps did most to reignite the old controversy and drag MR back into the philosophical limelight. Bechtel and Mundale express what seems to be an important insight about multiple realization: that unless we keep a consistent grain across realized and realizing kinds, claims alleging the multiple realization of psychological kinds are vulnerable to refutation. In this paper I argue that, intuitions notwithstanding, the terms in which their recommendation has been put make it impossible to follow, while also misleadingly insinuating that meeting their desideratum virtually guarantees mind-brain identity. Instead of a matching of grains, what MR really requires is a principled method for adjudicating upon differences between tokens. Shapiro's (2000) work on MR can be understood as an attempt to adumbrate such a method.

2. Bechtel and Mundale's grain requirement

Bechtel and Mundale appeal to "neurobiological and cognitive neuroscience practice" in the hope of showing how claims that psychological states are multiply realized are unjustified. Intuitively, theirs is an argument from success: cognitive neuroscience's method assumes MR is false, and the success of that method is evidence that MR *is* false. They argue that it is "precisely on the basis of working assumptions about commonalities in brains across individuals and species that neurobiologists and cognitive neuroscientists have discovered clues to the information processing being performed" (1999, 177).

Bechtel and Mundale examine both the "neuroanatomical and neurophysiological practice of carving up the brain." What they believe this examination reveals is, firstly, that the principle of psychological function plays an essential role in both disciplines, and secondly, that "the cartographic project itself is frequently carried out comparatively—across species" (1999, 177), the opposite of what one would expect if MR were "a serious option." It is the very similarity (or homology) of brain structure which permits generalization across species; and similarity in the functional characterization of homologous brain regions across

species only makes sense if the claims of MR are either false or greatly exaggerated. For instance, "[e]ven with the advent of neuroimaging, permitting localization of processing areas in humans, research on brain visual areas remains fundamentally dependent on monkey research…" (1999, 195). "The clear assumption is that the neural organization in the macaque will provide a defeasible guide to the human brain" (1999, 183). Brodmann's famous brain maps were based upon comparisons of altogether 55 species and 11 orders of mammals. If MR were true, "one would not expect results based on comparative neuroanatomical and neurophysiological studies to be particularly useful in developing functional accounts of human psychological processing" (1999, 178). They also argue that the ubiquity of brain mapping as a way of decomposing cognitive function points to the implausibility of the MR thesis. The understanding of psychological function is increasingly "being fostered by appeal to the brain and its organization" (1999, 191), again, the opposite of what one would expect "[i]f the taxonomies of brain states and psychological states were as independent of each other as the [MR] argument suggests" (1999, 190-91).

In light of such considerations, Bechtel and Mundale (1999, 178-79, 201-04) resort to grains as a way of making sense of what they perceive to be the

entrenched, almost unquestioning consensus prevailing around MR. They think that it can be traced to the practice of philosophers appealing to different grain sizes in the taxonomies of psychological and brain states, "using a coarse grain in lumping together psychological states and a fine grain in splitting brain states." When Putnam went about collecting his various specimens of pain, he ignored the many likely nuances between them. At the same time, he had few compuctions about declaring them different at a neurological level. His contention that pain is likely to be an MR kind can only command our respect if we can be sure that when he was comparing his specimens from a neurological point of view he was careful to apply no less lenient a standard of differentiation than he applied when comparing his specimens from a psychological point of view. Bechtel and Mundale maintain that when "a common grain size is insisted on, as it is in scientific practice, the plausibility of multiple realizability evaporates." As their examples of neuroanatomical and neurophysiological practice attest, scientists in these fields typically match a coarse-grained conception of psychological states with an equally coarse-grained conception of brain states. Despite the habit of philosophers individuating brain states in accordance with physical and chemical criteria, a habit no doubt originating with Putnam, this is not how neuroscientists characterize them. The notion of a brain state is "a philosopher's fiction" (1999,

177) given that the notion neuroscientists actually employ is much less fine-grained, namely "activity in the same brain part or conglomerate of parts."

A not unrelated factor is that the MR hypothesis often gets presented in a "contextual vacuum." The choice of grain is always determined by context, with "different contexts for constructing taxonomies" resulting in "different grain sizes for both psychology and neuroscience." The development of evolutionary perspectives, for instance, in which the researcher necessarily adopts a coarse grain, contrasts with the much finer grain that will be appropriate when assessing differences among conspecifics:

> One can adopt either a coarse or a fine grain, but as long as one uses a comparable grain on both the brain and mind side, the mapping between them will be correspondingly systematic. For example, one can adopt a relatively coarse grain, equating psychological states over different individuals or across species. If one employs the same grain, though, one will equate activity in brain areas across species, and one-to-one mapping is preserved (though perhaps further taxonomic refinement and/or delineation may be required). Conversely, one can adopt a very fine grain,

> and differentiate psychological states between individuals, or even in the same individual over time. If one similarly adopts a fine grain in analyzing the brain, then one is likely to map the psychological differences onto brain differences, and brain differences onto psychological differences. (1999, 202)

At least among some philosophers Bechtel and Mundale's message has evidently been well received (Couch 2004; Polger 2009; Godfrey-Smith, personal communication; see also tacit approval in Aizawa and Gillett 2009, 573). Polger (2009) explains the motivation for the grain requirement in an illuminating way. Neuroplasticity has in recent times been thought to provide compelling evidence for the MR of mental states. He concludes that "contrary to philosophical consensus, the identity theory does not blatantly fly in the face of what is known about the correlations between psychological and neural processing" (2009, 470). The grains argument figures prominently in his reasoning. As he points out, it might be tempting to regard a phenomenon like cortical map plasticity—where different brain regions subserve the same function at different times in an individual's history, say, after brain injury or trauma—as an existence proof of MR. But not if the point about grains is taken to heart. It all comes down to what we mean by "*different* brain regions" subserving "the *same* function." Consider that

recovered functions are frequently suboptimal. Genuine MR would indeed require the *same* psychological state to be underwritten by different neurological states; but suboptimality is evidence of difference underlying difference, not difference underlying sameness, as MR requires:

> It's true that this kind of representational plasticity involves the "same" function being mediated by "different" cortical areas. But here one faces the challenge leveled by Bechtel and Mundale's charge that defenses of [MR] employ a mismatch in the granularity of psychological and neuroscientific kinds. If we individuate psychological processes quite coarsely—by gross function, say—then we can say that functions or psychological states are of the same kind through plastic change over time. And if we individuate neuroscientific kinds quite finely—by precise cortical location, or particular neurons—then we can say that cortical map plasticity involves different neuronal kinds. But this is clearly a mug's game. What we want to know is not whether there is some way or other of counting mental states and brain states that can be used to distinguish them—no doubt there are many. The question is whether the sciences of psychology and neuroscience give us any way of *registering the two taxonomic systems*. (2009, 467, my emphasis)

3. Problems with the grain requirement: imprecise, impracticable, and misleading

But now the question is this: what, precisely, can it mean to use a "comparable" grain, or to keep a grain size "constant," across both psychological and neurophysiological taxonomies? Polger's motivation makes a lot of sense, to be sure, but talk of "registering" taxonomies (as of *aligning* classificatory regimes, or rendering distinct scientific descriptions *commensurable*, or however else one might care to put it) doesn't shed any light on how the desideratum for consistent grains can actually be met. Since it is intended to serve in part as a methodological prescription, it's important to know what to make of this requirement—metaphors won't help us here. How, in *concrete* terms, is an investigator meant to satisfy such a condition as *this* on their research?

Perhaps it means this. Suppose you have two tokens of fruit. The science of botany (say) could deliver descriptions under which the two are classified the same (e.g. from the point of view of *species*), but also descriptions under which they come out as different (e.g. from the point of view of *varieties*). The first

description could be said to apply a coarser grain than the second. Now imagine economics coming into the picture. The science of economics can likewise deliver descriptions under which both tokens are classified the same (e.g. both are forms of tradable fresh produce) or different (e.g. one, being typically the crunchier and sweeter variety, has a lower elasticity of demand than the other). Once again, the first description could be said to apply a coarser grain than the second. Perhaps, then, we could take it that botany and economics deliver descriptions at the same grain of analysis when their judgments of sameness or difference cohere in a given case. In the example, botanical descriptions via species classification would be furnished at the same grain as economic descriptions via commodity classification, so that species descriptions in botany are "at the same grain" as commodity descriptions in economics. By the same logic, *variety* descriptions in botany would be comparable to *elasticity* descriptions in economics. Fine. But if that is all that "maintain a comparable grain" amounts to, it really does beg the question, for this is simply type-type identity by fiat. *Of course* such a recommendation will ensure that the mapping between psychology and neuroscience will be "systematic" (to use Bechtel and Mundale's term), because on this account yielding concordant judgments of similarity or difference across taxonomies is what it *means* to apply the same grain. So we haven't solved the problem: *this* version of the grain

requirement makes type-type identity a fait accompli, effectively obliterating all MR kinds from the natural order.

It's just as well that I don't think this is what Bechtel and Mundale had in mind when they made their move to grains; supposing otherwise would serve only to trivialize an important aspect of their analysis. Still the construal is by no means far-fetched: "[o]ne can adopt either a coarse or a fine grain," they tell us, "but as long as one uses a comparable grain on both the brain and mind side, the mapping between them will be correspondingly systematic" (note that—it *will* be!). This sounds like someone with the utmost confidence in the grain requirement, which is of course what one *would* have if one thought grains could be legitimately matched in just this way. My guess is that, while they do have something important to tell us about MR, a beguiling metaphor has led them to suppose that MR is easier to refute than it actually is. (I'll support this contention with a few examples in a moment.)

Of course matters aren't much helped by the reasonable suspicion that MR is the result of pairing *inconsistent* grains. For what is neuroscience if not a fine-grained description of psychology, and psychology if not a coarse-grained

description of neuroscience? It is surely plausible that the neural and psychological sciences line up in something like this way, given that talk about the mind is really talk about the brain from a somewhat more abstract point of view.

What Bechtel and Mundale are ultimately trying to convey through their discussion of grains is the thought that claims of MR cannot be advanced willy-nilly—that there is an objective and standard way to go about verifying the existence of MR kinds and arbitrating disputes involving them. For the reasons just canvassed, however, it strikes me that talk of grains doesn't serve their purposes at all well. In fact they would have been nearer the mark had they said that what MR requires is some sort of principled *mismatching* of grains.

So far I've tried to indicate in what respects Bechtel and Mundale's grain requirement is imprecise and impracticable. Before I can show that the grains strategy is also misleading, and indeed often gets things wrong, I need to set it against an account which demonstrably gets things right.[1] Shapiro (2000) expresses with enviable lucidity what I think is the crucial insight towards which Bechtel

1 It is an account which even its detractors concede gets at least the essential point of interest to us here right, e.g. Gillett (2003).

and Mundale were uneasily groping. Interestingly, some philosophers—e.g. Polger (2009)—write as if the grain requirement and Shapiro's own formula for MR were effectively interchangeable. This is a mistake: the two approaches deliver different judgments in nontrivial cases (as I'll illustrate in a moment).

As Shapiro reminds us:

> Before it is possible to evaluate the force of [the MR thesis] in arguments against reductionism, we must be in a position to say with assurance what the satisfaction conditions for [the MR thesis] actually are. (2000, 636)

For him, "[t]he general lesson is this. Showing that a kind is multiply realizable, or that two realizations of a kind are in fact distinct, requires some work" (2000, 645). Furthermore, "[t]o establish [the MR thesis], one must show that the differences among purported realizations are causally relevant differences" (2000, 646). Shapiro's concerns revolve around what motivates ascriptions of difference, and therefore sameness. The issue is important because the classic intuition pump that asks us to conceive a mind in which every neuron has been replaced by a silicon chip depends on our ascription of an interesting difference between neurons and

silicon chips, apparently even where silicon chips can be made that contribute to psychological capacity by one and the same process of electrical transmission. His answer too, like Bechtel and Mundale's, depends ultimately on context—in particular, the context set by the very inquiry into MR itself.

Shapiro (2000, 643-44) argues that "the things for which [the MR thesis] has a chance of being true" are all "defined by reference to their purpose or capacity or contribution to some end." This is the reason why carburetors, mousetraps, computers and minds are standard fare in the literature of MR. They are defined "in virtue of what they do," unlike, say, water, which is typically defined by what it is, i.e. its constitution or molecular structure, and accordingly *not* an MR kind. Genuine MR requires that there be "*different* ways to bring about the function that defines the kind." Truly distinct (indeed *multiple*) realizations are those that "differ in causally relevant properties—in properties that make a difference to how [the realizations] contribute to the capacity under investigation." Two corkscrews differing only in color are not distinct realizations of a corkscrew, because color "makes no difference to their performance as a corkscrew." Similarly, the difference between steel and aluminium is not enough to make two corkscrews that are alike in all other respects two different realizations of a corkscrew "because, relative to

the properties that make them suitable for removing corks, they are identical." In this instance, differences of composition can be "screened off." Naturally there may be cases where differences of composition *will* be causally relevant (and it turns out that this will be important to the broader point I make below about where the grains strategy goes wrong). Perhaps rigidity is the allegedly MR kind in question. In that event, compositional differences will necessarily speak to how aluminium and steel achieve this disposition. The crucial thing to note here is that MR *is* the context, and MR makes *function* the relevant consideration, i.e. the specific point of view from which we will compare a set of tokens in the first instance (not phenomenology, not behavioral ecology, or anything else for that matter). Explanatory considerations may of course fine-tune the *sort* of function that captures our attention (cork-removal, rigidity, vision, camera vision, etc.). But function here is our key preoccupation, and having settled on a specific function which a set of tokens can be said to perform, the all-important question on Shapiro's analysis is *how* the two tokens bring that function about. Each case must be judged on its own merits. Thus unlike the two corkscrews identical in all respects save color, which do not count as distinct realizations, waiter's corkscrews and winged corkscrews are enabled to perform the same task in virtue of *different* causally relevant properties, and therefore *do* count as genuinely distinct

realizations of a corkscrew, one based on the principle of simple leverage, the other relying on a rack and pinions (Fig. 1).

(a) **(b)**

Figure 1. A waiter's corkscrew (a) and a winged corkscrew (b). Each contributes to the capacity of cork-removal in different ways.

Notice that to the extent Shapiro's causal relevance criterion envisages certain realizing properties being "screened off" from consideration in the course of inquiry, there is a sense in which the taxonomies of realized and realizing kinds may be said to be "commensurable" or "registrable" (no doubt explaining why some philosophers have simply confused commensurability with causal relevance). Thus when comparing the cork-removing properties of two waiter's corkscrews, compositional differences will not feature in the realizing taxonomy (if we accept Shapiro's characterization of the problem). So we have *cork-removal*,

which features in what we may regard as a coarse-grained taxonomy, realized by two objects described by a "science" of cork-removal in which microstructural variations do not matter, hence which might also be regarded as a coarse-grained taxonomy. If on the other hand we were comparing the same corkscrews for rigidity, where one was made of steel and the other of aluminium, compositional differences *would* feature in the realizing taxonomy. Here we would have *rigidity*, which features in what we could well regard as a more fine-grained taxonomy than that encompassing cork-removal, realized by two objects described by a science in which microstructural variations really *do* matter (namely metallurgy), and which might also be regarded as a fine-grained taxonomy, at least more fine-grained than the fictitious science of cork-removal. But my point is this: commensurability nowhere appears as an independent criterion of validity in Shapiro's account of MR, for it is an artifact of the causal relevance criterion, not a self-standing principle. Taxonomic commensurability is in fact an *implicit* requirement of the causal relevance criterion in the sense that it's taken care of once the proper question is posed. As an explicit constraint it is a will-o'-the-wisp.

Armed with this analysis, let's examine how Bechtel and Mundale attempt to refute the status of hunger as an MR kind. Putnam (1967) had compared hunger

across species as diverse as humans and octopuses to illustrate the likelihood that some psychological predicates are multiply realizable. On the basis of their grains critique, however, Bechtel and Mundale suggest that hunger will not do the work Putnam had cut out for it; for "at anything less than a very abstract level," hunger is different in octopuses and humans (1999, 202). The thought is that a finer individuation of hunger refutes the existence of a *single* psychological kind, hunger, which can be said to cross-classify humans and octopuses. Thus they essay to challenge the cognitive uniformity which MR requires at the level of psychology.

Perhaps we might first note that when identifying a *single* psychological state to establish the necessary conditions for MR, nothing Bechtel and Mundale say actually *precludes* the choice to go abstract. If context is what fixes the choice of grain (as they are surely right to point out), who's to say that context couldn't fix the sort of grain that makes hunger relevant in an abstract sense? It may be tempting to think that a more detailed description of something is somehow more *real*. But there is of course nothing intrinsically more or less real about a chosen schema relative to others that might have been chosen. There is no reason to suspect, for instance, that a determinate has any more reality than a determinable.

And yet there is a deeper problem with Bechtel and Mundale's deployment of the grains strategy here. To repeat their complaint: "at anything less than a very abstract level," hunger is different in octopuses and humans. But now why should *this* be relevant? Who would deny it? They themselves seem to be oblivious to the context which the very inquiry into MR makes paramount. They are not right to allege, as they do, that "the assertion that what we broadly call 'hunger' is the same psychological state when instanced in humans and octopi has apparently been widely and easily accepted without specifying the context for judging sameness" (1999, 203). The reason why hunger, pain, vision and so on were all taken for granted—assumed to be uniform at the cognitive level—is because MR made *function* the point of view from which tokens were to be compared. As Shapiro reminds us, "the things for which [the MR thesis] has a chance of being true" are all "defined by reference to their purpose or capacity or contribution to some end." It was understood that, say in the case of pain, regardless of phenomenal, ecological or behavioral differences between human and octopus pain (I doubt any of which were lost on Putnam), all instances of pain in these creatures had something like *detection and avoidance* in common. This might be to cast pain at "a very abstract level," but this just happens to be the context which

the inquiry into MR itself sets. A similarly abstract feature is what unites all instances of hunger: let's call it *nutrition-induction*. It is not that decades of philosophers had simply forgotten to specify the point of view from which these psychological predicates were being considered: it is rather that they simply didn't need to, since all of them had read enough of Putnam and the early functionalists to know what they were about. Phenomenal and other differences that one might care to enumerate between these predicates come a dime a dozen. But the whole point of functionalism was to abjure the inquiry into essences and focus instead on the causal role of a mental state within the life of an organism. Yes, this is to compare tokens from an "abstract level," but that's what made functionalism intriguing to begin with. And if Shapiro's analysis is any guide, it is really the *next* step in the endeavor to verify the existence of an MR kind that is the crucial one. Genuine MR requires that there be "*different* ways to bring about the function that defines the kind." So the follow-up question concerns *how* the relevant organisms achieve their detection and avoidance function, or nutrition-induction function, or whatever the case may be. It is in fact only by asking this next question that we can appreciate just how badly the grains strategy fares. The attempt to individuate hunger more finely does *not* refute the multiple realizability of hunger as between humans and octopuses. For, relative to the shared function of nutrition-induction,

it is extremely likely that humans and octopuses realize this capacity in different ways. The attempt to individuate pain more finely would likewise *not* refute the multiple realizability of pain as between humans and octopuses. For, relative to the shared function of detection and avoidance, it is extremely likely that humans and octopuses realize this capacity in different ways. So we see that the grains strategy, to the extent that it involves fine-graining psychological states in order to undermine the cognitive uniformity required by MR, sets itself a very easy job indeed, and mischaracterizes the nature of MR by its neglect of function. Moreover Shapiro's causal relevance criterion—which honors the core concerns motivating Bechtel and Mundale's resort to grains—does *not* demonstrate that hunger (or pain) is type-reducible.

A good illustration of the grains strategy in action is provided by Couch's (2004) attempt to refute the claim that the human eye and the octopus eye are distinct realizations of the kind *eye*. Conceding differences at a neurobiological level, the strategy again involves challenging the alleged uniformity at the cognitive level. As he explains, "[e]stablishing [MR] requires showing that…the physical state types in question are distinct [and] that the relevant functional properties are type identical. Claims about [MR] can be challenged at either step"

(2004, 202). Reminding us that psychological states "are often only superficially similar," and that "at a detailed level the neural differences make for functional differences" (2004, 203), he states:

> Psychologists sometimes talk about humans and species like octopi sharing the same psychological states. However, they also recognize that there are important differences involved depending on how finely one identifies the relevant features...Establishing multiple realization requires showing that the same psychological state has diverse realizations. But we can always disagree with the functional taxonomy, and claim there are psychological differences at another level of description. (2004, 203)

Thus he relates that while the two types of eyes have similar structure in certain respects, both consisting of a spherical shell, lens and retina, they use different kinds of visual pigments in their photoreceptors, as well as having different numbers of them, the octopus having one in contrast to the human eye which has four. They also have different retinas. The human retina, with rods and cones, focuses light by bending the lens and so changing its shape. The octopus eye, with rhabdomeres instead of rods and cones, focuses light by moving the lens

backwards and forwards within the shell. All these factors show up as differences in output, not just structure. The octopus, having only a single pigment, is colorblind, while its receptor's unique structure allows it to perceive the plane of polarized light. Retinal differences likewise make for functional differences, with very little information processing occurring on the octopus's retina, unlike the case of the human retina. This produces differences in stimuli and reaction times. So the two eyes might be similar, but when described with a suitably fine grain, he contends, they come out type distinct. In the result they are both physically *and* cognitively diverse, and so not genuine examples of MR.

Notice again that, contrary to what is claimed, it has not been demonstrated that type-type identity prevails here after at all (on the understanding that the kind camera $\text{eye}_{\text{human}}$ reduces to *its* distinct neural type, and the kind camera $\text{eye}_{\text{mollusc}}$ in turn reduces to *its* distinct neural type). If anything what this foray into mollusc visual physiology succeeds in showing is that, relative to the kind camera eye, human camera eyes and octopus camera eyes count as distinct realizations(!), for, assuming Shapiro's causal relevance criterion applies, human camera eyes achieve the function of *camera vision* differently to the way octopus camera eyes

achieve this function. Were we to attend to the original inquiry, which concerned whether human eyes and octopus eyes count as distinct realizations of the kind eye, Shapiro's own response, for what it's worth, is clear (2000, 645-46): here we do seem to confront a genuine case of type-type identity, as Putnam himself assumed, because, relative to the function of *vision* (not *camera vision*), both humans and molluscs achieve the function the same way (namely, by camera vision!). Differences that would be relevant at the neural level between humans and molluscs when asking how camera vision is achieved can be conveniently screened off when the question is how vision, as distinct from camera vision, is achieved. Again if pain or hunger were the kind in question, it seems more likely than not that we *would* confront a case of MR (unlike with vision), as we conjectured earlier. Explanatory context dictates the function of interest, and the function is one that we have to assume is common to the tokens in question in order to get the inquiry into MR off the ground. Indeed if Shapiro's analysis is correct, with MR we're always asking how some common function is achieved by different tokens that *do that thing*. Where there is no common function the question of MR cannot so much as arise. The fact that the question *does* arise in all the cases we've considered is a powerful indication that we're dealing with functions which all the relevant tokens actually share. The grains strategy confuses matters by suggesting that in many

cases involving putative MR kinds, psychological states can be individuated using a finer grain of description. But if what I have been saying is right, this is not the proper way to refute a putative case of MR.

That mine is the correct assessment of the situation is not only attested to by Shapiro's analysis of MR, but also by the fact that it avoids the very mug's game Polger sought to eschew by embracing the grains strategy in the first place. If for any putative MR kind I am free to cavil with the choice of your size of grain ("oh, that's far too coarse for psychology," or "now that's really not coarse enough for neuroscience"), how is the resulting game any less of a mug's game than the one we were trapped in at the start? I myself have played a few of these games with philosophers. No one wins. Couch's remarks are telling: "we can always disagree with the functional taxonomy, and claim there are psychological differences at another level of description." So the game goes on.

4. Conclusion

In sum, I think there's a genuine problem with the grain requirement. The central difficulty is that in the terms in which it's been put it is largely unworkable, and at

best no more than a loose metaphor. For a recommendation intended to serve at least in part as a methodological reform, this is clearly unsatisfactory. I don't deny that Bechtel and Mundale were onto something. But whatever value their insight into MR might have has been obscured by their unfortunate formulation of the issue. Moreover, as I have tried to show, the formulation is unfortunate not *just* because it happens to be unworkable. More worryingly, the argument from grains distorts the truth about MR by encouraging the view that mind-brain identity comes for free once we invoke the "same grain" of description across both realized and realizing kinds. But when the insight to which this locution seems to point is expressed in terms that are intelligible and empirically tractable (namely, Shapiro's causal relevance criterion), mind-brain identity seems anything but a fait accompli. Grains talk makes it tempting to think MR is easier to refute than it in fact is. It is certainly true, as Bechtel and Mundale acknowledge, that context fixes the choice of grain (where by "grain" we mean the respect under which we seek to compare a set of tokens); but we are not ipso facto obliged to employ a consistent grain across realized and realizing kinds (since this is just about meaningless as far as a researcher into these matters would be concerned and raises a host of difficulties beside). Rather than matching grains, what MR really behooves us to do is to apply a principled method for adjudicating upon differences between tokens of a

functional kind. Shapiro's work on MR shows us how to approach this important task.

References

Aizawa, Kenneth, and Carl Gillett. 2009. "Levels, Individual Variation, and Massive Multiple Realization in Neurobiology." In *The Oxford Handbook of Philosophy and Neuroscience*, ed. John Bickle, 539-81. New York: Oxford University Press.

Bechtel, William, and Jennifer Mundale. 1999. "Multiple Realizability Revisited: Linking Cognitive and Neural States." *Philosophy of Science* 66(2): 175-207.

Couch, Mark B. 2004. "A Defense of Bechtel and Mundale." *Philosophy of Science* 71(2): 198-204.

Gillett, Carl. 2003. "The metaphysics of realization, multiple realizability, and the special sciences." *Journal of Philosophy* 100(11): 591-603.

Polger, Thomas W. 2009. Evaluating the evidence for multiple realization. *Synthese* 167(3): 457-472.

Putnam, Hilary. 1967. Psychological predicates. In: *Art, mind, and religion*, eds. W. Capitan & D. Merrill, 37-48. Pittsburgh: University of Pittsburgh Press.

Shapiro, Lawrence A. 2000. "Multiple Realizations." *Journal of Philosophy* 97(12): 635-54.

Atlanta, GA; 3-5 November 2016 -605-

Interventionist Causation in Thermodynamics

Karen R. Zwier

March 2016 (Preprint)

Abstract

The interventionist account of causation has been largely dismissed as a serious candidate for application in physics. This dismissal is related to the problematic assumption that physical causation is entirely a matter of dynamical evolution. In this paper, I offer a fresh look at the interventionist account of causation and its applicability to thermodynamics. I argue that the interventionist account of causation is the account of causation which most appropriately characterizes the theoretical structure and phenomenal behavior of thermodynamics.

1 Introduction

The interventionist account of causation has been largely dismissed as a serious candidate for application in physics. For example, a dismissal of this sort is evident in the words of theoretical physicist Peter Havas:

> We are all familiar with the everyday usage of the words "cause" and "effect"; it frequently implies the interference by an outside agent (whether human or not), the "cause", with a system, which then experiences the "effect" of this interference. When we talk of the principle of causality in physics, however, we usually do not think of specific cause-effect relations or of deliberate intervention in a system, but in terms of theories which allow (at least in principle) the calculation of the future state of the system under consideration from data specified at a time t_0 (Havas 1974, 24).

And worries about the relevance of the interventionist account of causation in physics come not only from physicists, but also from philosophers—even those who favor interventionism:

> There are important differences between, on the one hand, the [interventionist] way in which causal notions figure in common sense and the special sciences and the empirical assumptions that underlie their application and, on the other hand, the ways in which these notions figure in physics (Woodward 2007, 67).

The reasons for dismissals and worries like those above are related to a common (but problematic) assumption that causation in physics has something to do with the dynamical evolution of a closed system. The problem is that, in our preoccupation with dynamical evolution and closed systems, we tend to forget and/or neglect those areas of physics for which we do *not* have complete equations of motion or for which it *doesn't make sense* to consider entirely closed systems. And it is in those areas that the dynamical view of physical causation makes less sense and interventionism finds its home.

In this paper, I propose to take a fresh look at the interventionist account of causation and its applicability to one of those neglected areas of physics: thermodynamics. I will argue that an interventionist analysis of thermodynamics succeeds where the dynamical view of physical causation fails. As I will show, all theorizing in thermodynamics requires careful definition of the "system" under consideration, which necessarily involves attending to the boundaries that enclose the system and the conditions imposed on those boundaries. Once boundaries are adequately specified, we end up with a strong distinction between the *internal* properties and processes of the system and those *external* influences that constrain the internal dynamics. It is in the distinction between internal properties and external influences that the natural fit between the structure of thermodynamic theorizing and the interventionist account of causation becomes apparent.

The plan of this paper is as follows. In section 2, I show that interventionist reasoning is inseparable from the structural foundation of thermodynamic theory. In section 3, I show how "driving forces" and their conjugate fluxes provide a rich basis for meaningful interventionist causal claims in thermodynamics. In section 4, I use the success of interventionist causal analysis in thermodynamics to make some broader concluding remarks.

2 The centrality of manipulated equilibrium

Thermodynamic theorizing is structured around the characterization of equilibrium states and the processes by which systems move from one equilibrium state to another. But just what is a thermodynamic equilibrium state?

A thermodynamic equilibrium state is the state of a system that is *not* undergoing a change (thermal, mechanical, or chemical). However, an equilibrium state is not a spontaneous occurrence. Natural thermodynamic systems are in constant flux. They engage in all sorts of interactions: they transfer heat, push and pull on one another, change their volume, and chemically react. The very idea of a thermodynamic "system", which can only be defined by the location and/or nature of its boundaries, is in itself a theoretical concept that we impose on the world in order to do thermodynamic "bookkeeping" (Dill and Bromberg 2011, 93). In order for a thermodynamic system to achieve an equilibrium state, the system must have been allowed to relax for a sufficient amount of time without the disturbing external influences of uncontrolled contact with other systems. And such a condition requires boundaries that isolate it—or

otherwise control exchanges—from other systems. Often those boundaries are put in place artificially, by human intervention.

Consider, for example, the air in an ordinary room. If we define our thermodynamic system in relation to the walls and doors of the room, we can say that the system has a fixed volume. If no massive weather change is currently occurring, we can assume that the air pressure in the room is approximately constant (not by isolation, but by contact with an external system whose pressure is approximately constant). If some kind of air conditioning system is in place and has been running for some time, we can also say that the temperature of the room is approximately constant. We can say that most of the chemical reactions occurring in the room are in a steady state and that the concentrations of various gases are relatively uniform (except perhaps for some minor concentration gradients near any plants and/or people located in the room), with equal flow into and out of the room for each type of gas. Notice, now, that even this *almost*-equilibrium state requires artificial maintenance (the rigidity of walls, contact with an exterior reservoir supplying constant pressure, the continuous work of the air conditioner, *etc.*). Stricter equilibrium states require much more careful isolation and maintenance, and true equilibrium states (which only exist in theory) require idealized boundaries (*e.g.*, perfect thermal insulators, frictionless pistons, perfectly rigid containers, *etc.*).

There is something of a tension, however, in the way that we think about equilibrium states. On the one hand, equilibrium states are the product of external conditions imposed on a system. On the other hand, once we consider those external conditions as given, a system will *naturally* or *spontaneously* tend toward the equilibrium state allowed by the constraints. But that spontaneous or natural behavior cannot be conceived of without external constraints being placed on the system in question. To even conceive of an equilibrium state, we must ask about the conditions imposed on its boundaries. What kind of walls enclose it? Permeable, semi-permeable, impermeable? Rigid or flexible? Adiabatic or conducting? There is no such thing as an equilibrium state unless the boundaries of the system are well-defined.[1] And the conditions imposed on those boundaries constitute external interventions on the system; they effectively *set* various thermodynamic variables to take on certain values. For example, conducting walls that put a system in contact with a thermal reservoir are effectively a way of *intervening* on temperature. Likewise, a semi-permeable boundary is a way of selectively *intervening* on particle concentrations in the system. (I will return to the question of how to conceive of boundary conditions as interventions on thermodynamic variables below in section 3.)

Thus, thermodynamic equilibrium states are inherently manipulated states—manipulated to be so either by human design or by some other mechanism that effectively imposes equilibrium conditions by external intervention. And these external manipulations or interventions, which impose values on certain thermodynamic variables, are entirely consistent with the concept of an intervention

[1] In fact, a system with no defined boundaries or external constraints is effectively a universe, and its fate is something like the "heat death" discussed by Thomson, Helmholtz, and Rankine.

that has been developed by Woodward (2003) and others. According to the interventionist account of causation, an intervention directly forces a variable to take on (or remain fixed at) a certain value. Furthermore, Woodward's definition of an intervention makes no reference to human action, and thus any entity or structure playing the role of setting certain variable values or holding them fixed can fulfill the requirements for intervention. For example, a cell membrane is a structure that effectively *intervenes* to maintain a certain equilibrium internal to the cell, by keeping interior and exterior pressures equal and by maintaining certain chemical concentrations by only allowing for select passage into and out of the cell.

Now how do these manipulated equilibrium states figure into theorizing about thermodynamic processes? We begin by representing our system of interest by reference to a *thermodynamic configuration space*. The thermodynamic configuration space is the set of all possible equilibrium states of a system, where the coordinates of that space are a relatively small number of macroscopic thermodynamic variables and each point in the configuration space represents a distinct equilibrium state. For example, we might choose as coordinates the following parameters: internal energy (U), volume (V), and the particle numbers of the various species present (N_1, N_2, ..., N_i). Then the entropy function for our system, $S = S(U, V, N_1, \ldots, N_i)$, will define a hyper-surface within the configuration space (see figure 1).

With this thermodynamic configuration space and the hyper-surface defined by the entropy function in place, we can begin to theorize about any ordered sequence of states (call these $A, B, C, \ldots$) located on the hyper-surface. Notice that a curve drawn through this sequence of states looks something like a process (in fact, we call it a *quasi-static process*) in that it represents a series of changes undergone by the system. However, such a curve can be nothing like a real process, because real processes involve nonequilibrium states and the curve represents a system that remains in equilibrium along its entire length. Furthermore, the curve could never represent the *autonomous* trajectory of a system, since every state that makes up the path is an equilibrium state and no isolated system would move from one equilibrium state to another spontaneously. So in order to think about a quasi-static process as something like a process, we must think of a system being "led"—by a series of external interventions—through the succession of desired states via "hops". We effectively imagine the system being "corralled" through the sequence of equilibrium states. And by imagining the sequence of hops between states to be very small and carried out by very tiny interventions, we can approximate a smooth curve more and more closely (in fact, arbitrarily closely).[2]

In summary, the structural foundation of thermodynamic theory is the set of equilibrium states and the quasi-static "processes" that can be drawn like lines through the space of such states. As I have argued here, the very idea of an equilibrium state is not possible without reference to boundaries and the constraints that *set* the value of certain thermodynamic variables within those

[2]My discussion here closely follows that of Callen (1985, Ch. 4).

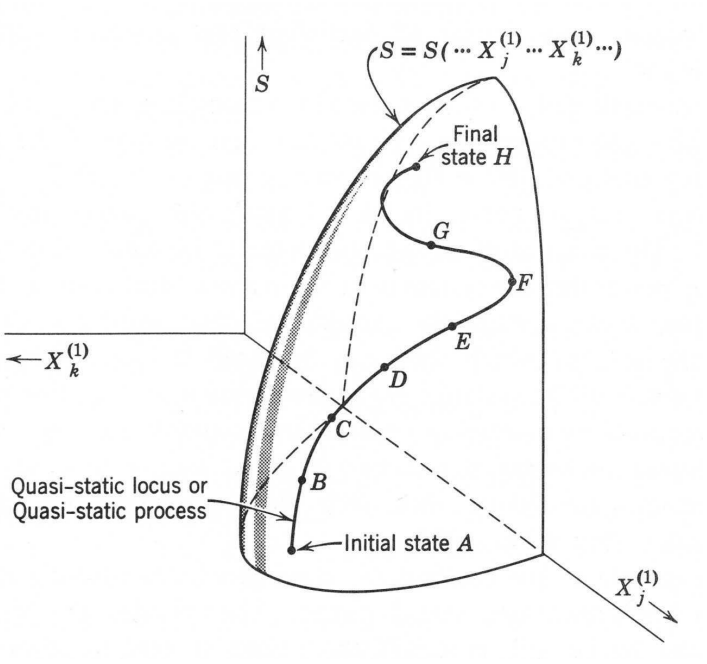


Figure 1: A representation of a quasi-static process in thermodynamic configuration space. From Callen (1985).

boundaries. Furthermore, we cannot think about quasi-static "processes", which are sequences of those equilibrium states, without thinking about a series of infinitesimal external interventions that force a system from one equilibrium state to the next. It is in this sense that interventionist reasoning is inseparable from the structural foundation of thermodynamic theory.

In the next section, I will discuss thermodynamic theorizing in greater specificity. As I will show, the interventionist view of causation maps naturally onto the use of potential functions when theorizing about a system undergoing a process.

3 Thermodynamic potentials and driving forces

The equilibrium state toward which a system will tend, given the conditions imposed on its boundaries, is governed by the energy and entropy considerations provided in the First and Second Laws of thermodynamics. The First Law tells us that any change in the internal energy (U) of a system will be equal to the total amount of energy it gains through energy exchange with the external world, in the form of heat and/or in the form of work. The Second Law tells us that any isolated system (*i.e.*, any closed system with fixed internal energy)

will tend toward its state of maximum entropy (S). The Second Law also has the result that the internal energy of any closed system with fixed entropy will be minimized. However, neither internal energy nor entropy are directly measurable, nor do we have a specific function that tells us their dependence on other state variables. What we do have, however, are other equations of state (*e.g.*, the ideal gas law) in addition to equations for U and S in *differential* form, which tell us about the way in which small changes in other state variables relate to small changes in energy and entropy:

$$dU = TdS - pdV + \sum_j \mu_j dN_j \tag{1}$$

$$dS = \left(\frac{1}{T}\right) dU + \left(\frac{p}{T}\right) dV - \sum_j \left(\frac{\mu_j}{T}\right) dN_j, \tag{2}$$

where T is absolute temperature, p is pressure, V is volume, μ_j is the chemical potential for species j, and N_j is the number of particles for species j. The above equations (and other variant forms) are commonly referred to as *thermodynamic potential functions*.

Notice that each term in both equations above involves a pair of conjugate variables. The second term in equation 1, for example, involves pressure and volume as a conjugate pair. For every pair of conjugate variables, one of the variables is extensive (*i.e.*, additive such that the property of a system is equal to the sum of that property for all of its component subsystems), while the other is intensive (*i.e.*, independent of the size of the system). Looking again at the second term in equation 1 as an example, pressure is the intensive variable and volume is the extensive variable.

Depending on the factors controlled in a given experimental context, each pair of conjugate variables tells us something about a tendency of the system as it moves toward equilibrium in that context. Since conjugate variables will be extremely important for our purposes here, let's concentrate on one pair and use an example to decipher its practical meaning.

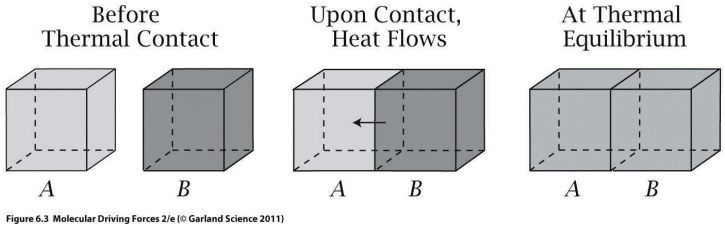


Figure 2: Two thermodynamic systems A and B before, during, and after arriving at thermal equilibrium. From Dill and Bromberg (2011, 100).

Consider the term $\left(\frac{1}{T}\right) dU$ in equation 2 and the process pictured in figure 2. We begin with two systems A and B, each enclosed in a rigid container. System A begins at temperature T_A and system B at T_B, where $T_A \neq T_B$.

The two systems are then brought into thermal contact with one another, but remain thermally insulated from the rest of the world. Now each system has an unknown entropy that can be expressed as a function of its internal energy, volume, and particle numbers, and since entropy is an extensive quantity, the total entropy of the combined system can be expressed as $S_{Total} = S_A(U_A, V_A, \mathbf{N}_A) + S_B(U_B, V_B, \mathbf{N}_B)$. Since entropy will be maximized at equilibrium, we use equation 2 to write the differential expression for S_{Total} and set it to zero:

$$\begin{aligned} dS_{Total} = \left(\frac{1}{T_A}\right) dU_A + \left(\frac{p_A}{T_A}\right) dV_A - \sum_i \left(\frac{\mu_{A_i}}{T_A}\right) dN_{A_i} + \\ \left(\frac{1}{T_B}\right) dU_B + \left(\frac{p_B}{T_B}\right) dV_B - \sum_j \left(\frac{\mu_{B_j}}{T_B}\right) dN_{B_j} = 0 \end{aligned} \tag{3}$$

If we assume that there is no particle exchange between the two systems and that no chemical change occurs within each system, we can eliminate the terms that allow for changing particle numbers. And since the containers are rigid, we can eliminate the terms that allow for changing volume. Furthermore, given that the combined system is isolated from the external world, the total internal energy of the combined system must remain constant, and any change in energy of either system must be compensated by a change in energy of the other. Thus, $dU_A = -dU_B$. So we have the following simplified expression:

$$dS_{Total} = \left(\frac{1}{T_A} - \frac{1}{T_B}\right) dU_A, \tag{4}$$

which will be equal to zero (*i.e.*, attain equilibrium) when $T_A = T_B$.

Thus we have derived the well-known result that two objects brought into thermal contact will reach equilibrium when their temperatures are equal. But more importantly for our purposes here, we can interpret the factors in equation 4 in light of this equilibration process. The difference in temperatures between the two systems leads to a nonzero value of the factor $\frac{1}{T_A} - \frac{1}{T_B}$, which effectively acts as a "force" driving a change dU_A in the internal energy of system A. More generally speaking, when a system is placed in thermal contact with a system at a different temperature, the temperature difference between the two systems acts as a force driving an exchange of heat energy between the systems. Phrased in terms of a system and its surroundings, $\frac{1}{T}$ describes the tendency of a system to exchange heat with its environment; it is the incremental relaxation that a system experiences in transferring a small bit of its energy dU.[3]

Physicists commonly use the language of "driving forces" in referring to the intensive parameters in the thermodynamic potential functions. Looking back again at equation 2, a difference between the pressure p of the system and its environment will act as a driving force for an exchange of volume dV between the system and its environment, and a difference between the concentration of a

[3] Alternatively, we could have begun with the thermodynamic potential function for internal energy (equation 1) to derive the same result.

particular species μ_j in the system and its environment will act as a driving force for exchanges of particles of the respective species with the environment (dN_j). The force or tendency represented in each of the conjugate pairs $(T, p, \boldsymbol{\mu})$ can act, separately or together (depending on the constraints imposed on the process), to drive changes in its paired extensive variable (dU, dV, or $d\mathbf{N}$, respectively), and thus to drive the system and its environment toward the equilibrium state of maximum entropy.[4]

This "driving force" language—and its basis in the way in which the environment exchanges energy and entropy with a system—matches the way in which relationships among thermodynamic variables would be modeled by the interventionist account of causation. According to the interventionist account, a variable X is an interventionist cause of another variable Y if there is a possible intervention on X that will change the value of Y (or the probability distribution over the values of Y) when the values of all other variables in the system remain fixed.[5] In physical experiments, the condition that the values of all other variables in the system remain fixed across changes in the intervention on X is often enforced using what I will call "auxiliary interventions" on those variables. To see how interventionist treatment matches the "driving force" language, let's consider the temperature equilibration case above, with system A as the causal system under investigation.

Consider the set of thermodynamic variables characterizing system A when we consider the temperature equilibration process in terms of maximization of entropy: volume V_A, the set of particle numbers for each species $\mathbf{N}_A$, temperature T_A, and internal energy U_A. Each of these variables is represented below in figure 3. The primary intervention in the temperature equilibration case was the operation of placing system B in thermal contact with system A. This intervention occurred specifically under conditions in which the volume V_A and particle numbers $\mathbf{N}_A$ of system A were held constant; the enforcement of constant values of V_A and $\mathbf{N}_A$, by enclosing the system within rigid impermeable walls, constitutes the set of auxiliary interventions in this case. Under the conditions set by these auxiliary interventions, the primary intervention produced a change in T_A, since the original temperatures of the two systems were not equal, and this change in temperature resulted in a change in the internal energy (U_A) of the system. And since, under conditions where all other variables are held constant, the intervention was an intervention on T_A and resulted in a change

[4] Physicists use the language of "driving forces" in both the entropy and energy representations. When we flip between the energy picture of a system and the entropy picture of that same system, the metric by which we measure progress toward equilibrium changes. Each metric has its own way of characterizing the driving force because, in changing our metric of progress, there is a transformation on the force term. Still, physically, it is one and the same force driving the system toward equilibrium. This representational change in the physical equations mirrors a widely-noted feature of the interventionist account of causation: when we change the set of variables with which we characterize a causal system, our characterization of the causal relationship itself can change.

[5] I have ignored some technical details for the sake of simplicity here. See Woodward (2003, 59) for the more precise interventionist criteria for X's being a type-level direct cause of Y and X's being a type-level contributing cause of Y.

in U_A, we can say that T_A is an interventionist cause of U_A.

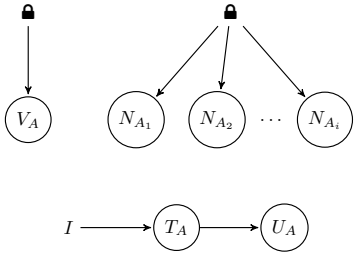


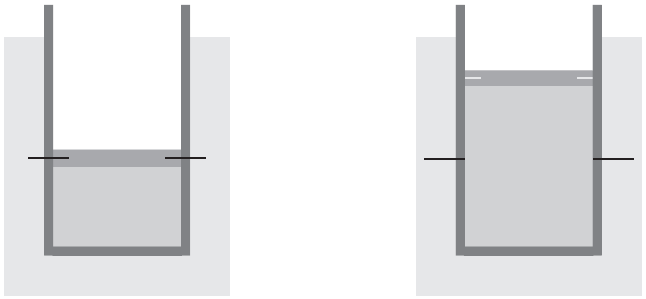
Figure 3: An interventionist causal graph of the temperature equilibration process in which system A, originally at temperature T_A, is brought into contact with another system B, originally at temperature T_B. The variable I represents the intervention that places the two systems in contact and thus changes the value of T_A. The lock symbols (🔒) represent the auxiliary interventions which hold V_A and $\mathbf{N}_A$ fixed.

To further flesh out the causal claim being represented by the arrow from T_A to U_A in figure 3, we can contrast varying interventions in which we put system A in contact with system B at varying temperatures $T_{B1}, T_{B2}, \ldots T_{Bn}$, while still holding V_A and $\mathbf{N}_A$ constant at the same values. Under such varying interventions, we will find that there are corresponding variations in the final T_A and U_A. Therefore, the interventionist account confirms that the temperature T_A of system A is a cause of its internal energy U_A. In general, interventions on temperature lead to changes in internal energy via exchange of heat when volume and particle numbers are held constant. Such a causal claim seems to be precisely what physicists mean to convey when they use "driving force" language with respect to temperature.

The intervention in the above case, where we have an equilibration process between two finite systems with differing initial temperatures, is an example of a "soft" or "parametric" intervention in that it *modifies* the temperature of our system rather than determining it completely.[6] When we put system A with its initial temperature T_A in contact with system B with its initial temperature T_B, the combined system finds an equilibrium temperature somewhere between the initial values of T_A and T_B. But thermodynamics also provides conceptual tools for theorizing about "hard" or "structural" interventions that entirely determine the value of an intensive parameter for a system. We call these theoretical entities "reservoirs" or "baths", and they have the property of being able to exchange one or more extensive quantities while their corresponding intensive properties remain constant. For example, an energy bath (*i.e.*, a temperature reservoir), by virtue of its size, is able to exchange energy with a system with which it is put in contact with negligible effect on its temperature. Likewise, a volume bath (*i.e.*, a pressure reservoir) is able to exchange volume while remaining at constant pressure, and a particle bath (*i.e.*, concentration reservoir) is able to exchange particles while maintaining constant particle con-

[6] For the distinction between soft and hard interventions, see Eberhardt and Scheines (2007).

centrations. When we theorize about cases in which we put a system in contact with a reservoir instead of a finite system, we consider a hard intervention that *determines* the value of the relevant intensive variable in our system. Such theoretical experiments bring the interventionist causal structure into even clearer relief: putting a system in contact with a reservoir is an intervention that sets the value of an intensive variable in the system, which in turn results in a change in the corresponding extensive variable.



(a) (b)

Figure 4: Illustration of a pressure-driven process, depicting (a) the system in its initial equilibrium state before the piston-locking pins are released; (b) the system once it has reached its new equilibrium state after the pins are released. This image shows the result of the case where $p_0 > p_{Res}$ and the piston rises, but all of the same considerations would apply in the case that $p_0 < p_{Res}$ and the piston falls.

Let's look at an example. Consider a system that is in an initial equilibrium state $(p_0, T, \mathbf{N})$. Suppose that we intervene on the system by bringing it into contact with a reservoir that maintains the same temperature T as the system but a different pressure p_{Res}. We might do so by releasing an initially-locked piston, allowing it to move freely between the system and the reservoir (see figure 4). The process that ensues will be ruled by a maximization of the entropy of the total combined system, so we are interested in the condition where $dS_{Total} = 0$:

$$dS_{Total} = \frac{1}{T_{Res}} dU_{Sys} + \frac{p_{Sys}}{T_{Res}} dV_{Sys} + \frac{1}{T_{Res}} dU_{Res} + \frac{p_{Res}}{T_{Res}} dV_{Res} = 0 \quad (5)$$

Due to conservation of volume and conservation of energy, $dU_{Sys} = -dU_{Res}$ and $dV_{Sys} = -dV_{Res}$, so the above condition reduces to the following:

$$dS_{Total} = \left(\frac{p_{Sys} - p_{Res}}{T_{Res}} \right) dV_{Sys} = 0 \quad (6)$$

We can see here that it is the pressure difference between system and reservoir that is driving the exchange of volume. And again, this physical interpretation in terms of driving forces matches the interventionist causal account. By placing the system in contact with the reservoir, we *set* the pressure of the system to a new value, and the forced change in pressure results in a change in volume. Were we to impose a different pressure on the system by placing it in contact

with a reservoir at a different pressure, we would see the corresponding volume change as well. Thus, pressure is an interventionist cause of volume (see figure 5).

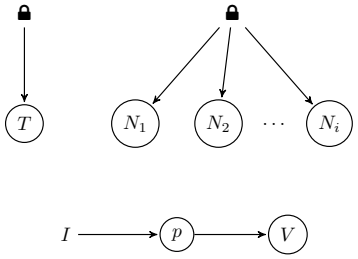


Figure 5: Interventionist causal representation of the pressure equilibration process depicted in figure 4. The variable I represents the intervention that places the system in contact with the pressure reservoir and thus changes the value of p. The lock symbols (🔒) represent the auxiliary interventions which hold T and $\mathbf{N}$ fixed.

As shown in the examples above, the most important key to successful thermodynamic theorizing is the careful definition of the boundaries between systems and accounting for the transactions that occur at those boundaries. Interventionist reasoning fits naturally into thermodynamic theorizing because its distinction between the interventions external to a causal system and the causal relations internal to that system is perfectly applicable where thermodynamic boundaries are well-defined. Since interventions are always performed *on* a causal system from outside, it is entirely natural to label exchanges between a system and its environment as interventions of the environment on those systems.

4 Conclusion

In this paper, I have shown that there is a natural fit between thermodynamic theorizing and the interventionist account of causation. I therefore argue that the interventionist account is the most suitable account of causation for describing thermodynamic theorizing and our actual interactions with thermodynamic systems.

I suggested at the beginning of this paper that we tend to assume that physical causation will have a dynamical form, and that my identification of interventionism as the most appropriate account of causation in thermodynamics would run contrary to this assumption. It might be objected that this is a somewhat dull result, however. Thermodynamics, so the objection might run, is not "fundamental" physics, and so it is unsurprising that we should find interventionist causation rather than dynamic causation in a realm of physics that is…well…*not dynamical*. But such an objection would miss the point. Our common assumption that "physical causation" must refer to the dynamical propagations of systems is the result of our preoccupation with "fundamental" physics (which

we also assume, almost by definition, must have a dynamical form) and neglect of those areas of physics which are considered to be "non-fundamental".[7]

So what is it to be a cause in (at least some of) physics? Here is a simple answer: an account of causation which appropriately characterizes the theoretical structure and phenomenal behavior of a domain of physics gives an account of what it is to be a cause in that domain of physics. And I have shown that the interventionist account does just that in thermodynamics.

References

Batterman, Robert, ed. 2013. *The Oxford Handbook of Philosophy of Physics*. Oxford: Oxford University Press.

Callen, Herbert B. 1985. *Thermodynamics and an Introduction to Thermostatics*. 2nd ed. New York: John Wiley & Sons.

Dill, Ken A. and Sarina Bromberg. 2011. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*. 2nd ed. New York: Garland Science.

Eberhardt, Frederick and Richard Scheines. 2007. "Interventions and Causal Inference." *Philosophy of Science* 74 (5): 981–995.

Havas, Peter. 1974. "Causality and Relativistic Dynamics." In *Causality and Physical Theories*, edited by William B. Rolnick, Vol. 16 of *AIP Conference Proceedings*, 23–47. New York: American Institute of Physics.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

——— 2007. "Causation with a Human Face." Chap. 4 in *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by Huw Price and Richard Corry, 66–105. Oxford: Clarendon Press.

[7] Increasingly, the study of "non-fundamental" theories has revealed that their relationship with "fundamental" theories is less straightforward than might be expected. For recent discussions in this vein, see, for example, Batterman (2013). Furthermore, it is entirely unclear what the criteria for "fundamental" status are, or whether undisputed criteria even exist. And with the criteria for fundamentality in doubt, it is hard to see what basis we might have for even expecting "fundamental" theory to always have dynamical form.

Anthropomorphism as Cognitive Bias

Mike Dacey

Philosophers and psychologists have long worried that a human tendency to anthropomorphize leads us to err in our understanding of nonhuman minds. This tendency, which I call *intuitive anthropomorphism*, is a heuristic used by our unconscious folk psychology to understand the behavior of nonhuman animals. I argue that the dominant understanding of intuitive anthropomorphism underestimates its complexity. It does often lead us to err, but not always. And the errors it produces are not only overestimations of nonhuman intelligence. If we want to understand and control intuitive anthropomorphism, we must treat is as a cognitive bias, and look to the empirical evidence. The literature on controlling implicit social biases is particularly helpful. That literature suggests that the most common for intuitive anthropomorphism, Morgan's Canon, should be rejected, while others need supplementation. It also suggests new approaches.

1. Introduction

Humans naturally anthropomorphize. As David Hume put it: "There is an universal tendency among mankind to conceive all beings like themselves . . . We find faces in the moon, armies in the clouds" (Hume 1957, pg. 29). Philosophers and psychologists attempting to understand the minds of nonhuman animals have long worried that this tendency leads us to error. This worry is shared across fields and across theoretical attitudes about anthropomorphism more generally. Kennedy (1992) argues forcefully against any form of anthropomorphism, and sees this tendency, which is "simply *built into us*" (pg. 28, emphasis his), as the reason it is so problematic. Rivas & Burghardt (2002) believe that some forms of anthropomorphism are valuable, but caution against its naïve forms: "Anthropomorphism is like Satan in the bible - it comes in many guises and can catch you unawares!" (pg. 15). I call the human tendency to anthropomorphise *intuitive anthropomorphism*, and it is the specific target of this paper. Existing views get intuitive anthropomorphism wrong, and as a result, fail to control its effect on the sciences of nonhuman minds.

To start, the term 'anthropomorphism' (*simpliciter*) needs clarification. In its strictest sense, anthropomorphism is sometimes defined as a *kind of error*: overestimating the intelligence of animals by attributing to them human-like traits they do not have (perhaps consciousness or belief-desire psychology). A broader sense of the term applies to the belief that a nonhuman animal possesses any of these 'characteristically human' traits, while allowing that that belief may be true or false. A still broader sense treats anthropomorphism as a way of thinking, or a *process* of forming beliefs about non-human minds: coming to understand the minds of nonhuman animals by analogy to our own, while leaving open whatever beliefs we might form. I treat anthropomorphism as a process. Though anthropomorphism-as-an-error is probably most often made explicit, usage of the term tends to slide between these, especially in debates over the role of anthropomorphism in science (more in section 6). I view anthropomorphism as a process because it allows more productive engagement with the crucial problem.[1] I say this for three reasons.

Firstly, this use allows me to remain agnostic about anthropomorphism generally. There is a large and contentious debate about whether there are legitimately scientific anthropomorphic strategies (e.g. Burghardt 1985, Rivas & Burghardt 2002, de Waal 1991, 1999, Mitchell 2005, Wynne 2007). I side-step this debate to focus specifically on intuitive anthropomorphism, which is just one kind of anthropomorphism.

Secondly, it does not prematurely restrict anthropomorphism to certain posits, better capturing the nature of intuitive anthropomorphism. Intuitive anthropomorphism is a heuristic

[1] There are two other reasons I won't argue for at length. First, treating anthropomorphism as a process is anthropocentric in a much less pernicious way than the other definitions. They mark off a class of 'characteristically human' traits, but why think we have special claim to them? But, we are human, so it is natural that we interpret other species from our own perspective. Second, this view shows how one can be concerned about intuitive anthropomorphism without thinking that animals are unintelligent, or that they don't have minds. This is not what is at issue here.

employed by our unconscious folk psychology. Our unconscious interprets behavior of nonhuman animals in the same way it interprets human behavior. It is an empirical question what effects this may have. Like any cognitive heuristic, it likely leads to errors in many cases, but it does not always do so, and we can't know how or when it does until we've tested it. It is a mistake to assume what effects it has at the outset. In fact, intuitive anthropomorphic error is *not* merely a matter of overestimating intelligence by positing a set of specific mental states, so controls that simply aim at correcting this mistake are ineffective. The effects of intuitive anthropomorphism are complex, and the errors it produces share nothing besides their common source. Focusing on any particular kind of posit or error before we understand intuitive anthropomorphism dooms us from the start.

Thirdly, this way of looking at anthropomorphism opens up a new approach to debates about it. Too often these debates proceed as follows: one theorist claims that common explanations of some behavior (read: not their own preferred theory) are either overly anthropomorphic or overly averse to anthropomorphism, while another claims the opposite. These are not really arguments about anthropomorphism itself; they are arguments about *those theories* of animal cognition. Very little in comparative psychology is settled, so it should not be surprising that these debates make little headway. This approach makes sense if anthropomorphism is a kind of error: to identify instances of anthropomorphism in the field, we must first identify errors in the field. If anthropomorphism is a process, there is another approach available at the level of individual psychology. So I do not argue that the field is overly anthropomorphic, I argue that any particular judgement one makes about the psychology of nonhuman animals might be subject to the influence of intuitive anthropomorphism. As such, any particular judgement is potentially subject

to an intuitive anthropomorphic bias.[2] To understand what this means, I look to the empirical literature.

Intuitive anthropomorphic bias is one kind of cognitive bias that results from reasoning heuristics applied in our unconscious (e.g. Tversky & Kahneman 1974). The literature on cognitive biases and unconscious processing generally will help understand intuitive anthropomorphism (section 2). The literature on implicit social biases will help understand how to control it (section 3). This is the largest literature on controlling unconscious biases, and it targets interventions of the right form for the current discussion. Collectively, this literature suggests that existing strategies for controlling intuitive anthropomorphism are ineffective (sections 4-6), and can help develop new controls (section 7).

Figure 1.

[2] This might appear to *imply* that the field is overly anthropomorphic. But I argue that the effects of intuitive anthropomorphism are poorly understood, so it is unclear what it means to say that the field is overly anthropomorphic in the first place.

2. The Intuitive Anthropomorphic Bias

Figure 1 shows Ham the Chimpanzee in 1961, on his way to the capsule that will launch him into space. He is about to become the first hominid to orbit the earth. The look on his face appears to be one of excitement or pride (internet comments on the photo show how common this interpretation is). Perhaps he perceives excitement in the behavior of those around him, or perhaps he is just pleased by the attention. Unfortunately, this is unlikely. In chimpanzees, this expression is known as the 'fear grin.' As much as we can rationalize the thought that Ham is pleased, fear is a more likely reactions to being strapped into a seat and carried to a strange enclosure.

This is a case where intuitive anthropomorphism leads us astray. We see an animal grinning and it *looks to us* like happiness. That is a pretty good heuristic for dealing with other humans, but we go wrong because our unconscious folk psychology applies it the same way in nonhumans. Even knowing that we are wrong, the perceptual Gestalt remains: Ham still looks happy. Note that we are not positing human-like capacities in Ham that he incapable of having. There is no substantive difference in the intelligence required for fear and happiness.[3] Intuitive anthropomorphism does lead us to err, and those errors do not just overestimate intelligence (though sometimes they do). We also should not think that intuitive anthropomorphism *always* leads us to error (anthropomorphism is a way of thinking, not a kind of error), but there is good reason to think it will do so often.

Daily experience with pets and animal cartoons demonstrates how easy it is: we talk to our dogs, and do not blink when cartoon dogs talk to their owners. There are also evolutionary

[3] I am speaking loosely of 'happiness' and 'fear,' but I do not mean the *human* mental states; I mean whatever is the chimpanzee analogue.

reasons to think we *should* anthropomorphize: Intuitive anthropomorphism is the kind of fast and frugal heuristic that can work for evolutionary purposes, even if it is not up to the epistemic standards of science. As some have argued (e.g. Caporeal & Heyes 1997, Gallup, Marion & Eddy 1997), our intuitive folk psychology most likely evolved to inform social interactions with other humans, and then was exapted to handle interactions with nonhuman animals. The intractable problems of interpreting other minds are a prime target for 'good enough' predictive strategies (Dennet 1989 argues in a similar spirit).

There is also considerable empirical evidence that humans make errors because of intuitive anthropomorphism. In one of the classic studies of psychology, Heider & Simmel (1944) showed that participants attributed intentional actions to a collection of two dimensional shapes 'interacting' in a short cartoon. This indicates a tendency to see (quite literally) intentional action when certain cues are present. In this experiment, the cues are irregular movements that seeming respond to one another. This, by itself, is insufficient evidence to really conclude that behavior is intentional, but it seems to be enough for our unconscious folk psychology.

In general, our anthropomorphic unconscious folk psychology seems to trigger on simple or irrelevant cues, suggesting it triggers too often. Other simple cues like hands, eyes, and faces influence the attribution of conscious mental states (e.g. Arico et. al. 2011, Fiala, Arico, & Nichols 2011). Magnifying this, humans are wired to err in the direction of seeing faces that aren't there over missing faces that are (e.g. Liu et. al. 2014). Finally, humans are biased in attributing conscious mental states to animals that move at about the same speed as us (Morewedge, Preston, & Wegner 2007).

Children, of course, anthropomorphise wildly (e.g. Gebhard, Nevers, & Billmann-Mahecha 2003). And adults anthropomorphize when explaining and imagining behavior. Religious

participants asked to describe God or tell stories involving God will anthropomorphize God, even when doing so contradicts their theological beliefs (e.g. these are anthropomorphic errors by the subjects' own lights; Barret & Keil 1996). Participants presented with written descriptions of situations and asked to assess the 'reasonableness' of various mental state attributions ascribe mental states to dogs that are of the same kind as they do a human child (though quantitatively simpler; Rasmussen and Rajecki 1995).

One might wonder whether these effects apply to scientists. While scientists may not make the flagrant errors that children make, or that experimental subjects make in snap judgments, we should expect that they do make some errors. In general, unconscious social biases influence behavior even during careful deliberation by experts. This has been shown in hiring decisions (Bertrand & Mullainathan 2003), including hiring by scientists (Moss-Racusin et al. 2012), medical decision making (Green et. al. 2007), the judicial process (Banks, Eberhardt, & Ross 2006, Mitchell et. al. 2005, Rachlinski et. al. 2009). Expertise and deliberation are not themselves sufficient to stop cognitive biases. In fact, the self-perception that one is objective *increases* social bias (Uhleman & Cohen 2007).

So scientists and philosophers should not expect themselves to be immune. In addition, scientific practice is complex, and intuitive anthropomorphism can arise at every stage of research: model construction, experimental design, data gathering, and model choice. At each stage, it can arise in different ways, and bias at one stage will be compounded in later stages: attention and recall are biased towards stereotype confirming evidence (e.g. confirmation bias supporting implicit bias; Bodenhausen & Wyer, 1985; Darley & Gross, 1983). This, along with the subtlety of cognitive biases generally, mean that the effects of anthropomorphic bias can be complex and, in at least some cases, counterintuitive (recall Ham).

Putting this all together, there is good reason to believe that intuitive anthropomorphism is problematic in a context like comparative psychology. It is triggered by simple stimuli like eyes and hands, and its effects can be subtle and unpredictable. Awareness of the bias, even with deliberation and expertise, is not sufficient to control it. We cannot be sure how and to what degree any specific judgment about animal cognition is (or is not) influenced. So what can we do?

3. Controlling Implicit Social Bias

The literature on controlling implicit biases provides the best current evidence about controlling cognitive biases, and some of the interventions discussed there are especially instructive to the current discussion. So I turn to that literature now. The upshot is that taking steps to ensure that counter-stereotypical information is salient in reasoning is more effective than simply directing participants to avoid or 'correct' for bias.

The Weapon Identification Task is a common test of bias. Participants identify an image as a tool or a weapon, but are very briefly shown an image of a white or black face before. If, for instance, subjects mistakenly say that a tool is a weapon after a black face more often than a white face, that is taken as a characteristic indicator of bias.

Not all interventions that make intuitive sense work. Using this task, Payne, Lambert, & Jacoby (2002) tested the effects of instructions on bias. Participants in the two experimental conditions were first told that research had shown that people possess implicit racial biases that can influence performance on the task, and then told either to "avoid race" or "use race" in making their decision. Controls were not given any of these instructions. In fact, the two experimental groups were *more likely* to make errors consistent with racial bias than the control

group. Thus, the authors conclude, calling attention to race has the effect of *increasing* bias, whether participants are told to use or avoid it. The instruction itself activates racial stereotypes that lead to biased judgments. Similarly, Legault, Gutsell, & Inzlicht (2011) showed that motivating participants to avoid racial bias *increases* bias if they perceive the motivation as coming from an external source, but can reduce bias if it is seen as self generated.

Stewart & Payne (2008) found a more effective approach. They used *implementation intentions*, which are if-then action plans that make it easier for participants to accomplish a goal than general intentions to do so. For instance, the implementation intention "if I leave work, I will stop and exercise at the gym" is more effective than the general intention "I will exercise more." Stewart & Payne asked participants to form one of three implementation intentions. The first: "whenever I see a black face on the screen, I will think the word, *accurate*." The second: "whenever I see a black face on the screen, I will think the word, *quick*." The third: "Whenever I see a black face on the screen, I will think the word, *safe*." Those participants were told: "By thinking the word 'safe,' you are reminding yourself on each trial that you are just as safe interacting with a Black individual as with a White individual" (pg. 1336). Only the 'think safe' condition reduced bias.

Another common intervention is to show participants images of admired or counter-stereotypical members of the stereotyped group. Dasgupta & Greenwald (2001) found this effect using the Implicit Association Test. In one manipulation, Dasgupta & Greenwald (2001) showed participants images of admired black individuals and disliked white individuals before giving them the IAT. They found that doing so reduced bias, whether the images were presented immediately before or 24 hours before administering the IAT. Govan & Williams (2004) achieved a similar result using counter-stereotypical examples in the experiment itself, even

using non-social stereotypes about flowers and insects. Imagining a positive, productive interaction with members of the stereotyped group before performing an IAT can also reduce bias (Turner & Crisp 2010).

So the most consistently effective interventions are those that make counterstereotypical information salient in reasoning, like the 'think safe' intention, or imagined interactions. What doesn't work is demanding accuracy. I now apply the lessons of the discussion so far to existing methods of controlling anthropomorphism.

4. Morgan's Canon

Perhaps the most commonly advocated method of addressing anthropomorphism is an updated version of *Morgan's Canon*, a famous statement by the 19th century comparative psychologist C. Lloyd Morgan (1894). The modern interpretation of Morgan's Canon councils that researchers should adopt the model that describes the *simplest* psychological process that can predict behavior. This practice is widespread, and is still motivated largely by concerns about anthropomorphism (Graham 1998, Manning & Dawkins 1998, Shettleworth 2010, Wilder 1996, Wynne 2007). In order to control anthropomorphism generally, it must control intuitive anthropomorphism specifically.

de Waal (1998), Sober (2005), and Fitzpatrick (2008) have argued that Morgan's Canon leads to errors of 'anthropodenial,' the underestimation of animal intelligence (more on their reply in section 5). My arguments so far show why: Morgan's Canon explicitly aims to correct a bias that *consistently overestimates* intelligence, but, first, intuitive anthropomorphism does not always lead to errors, and second (think of Ham) its errors need not have anything to do with

intelligence. So Morgan's Canon sometimes aims to correct errors that were not made, and sometimes aims to correct the wrong error.

My arguments also suggest another problem: even in cases where intuitive anthropomorphism has lead someone to overestimate the intelligence of an animal, Morgan's Canon may not be effective. Effective interventions to control implicit bias make counter-stereotypical information salient. Morgan's Canon does not. It is more similar to the ineffective interventions: it is a demand to make the 'unbiased' judgment, like the Stewart & Payne (2008) 'think accurate' condition, or the Payne, Lambert, & Jacoby (2002) 'avoid race' instruction. The effect Morgan's Canon has depends on how it is represented by the person using it. If researchers view it as an external demand to avoid anthropomorphism, it could actually *increase* bias (Legault, Gutsell, & Inzlicht 2011). Similarly, if researchers view it as being *about* anthropomorphism, it could activate anthropomorphic stereotypes, and increase bias, like the "use race" and "avoid race" instructions from Payne, Lambert, & Jacoby (2002).

So how this plays out will vary on a case by case basis. But this leaves Morgan's Canon in a bad position. Firstly, intuitive anthropomorphism does not always lead to error, and when it does, it does not always overestimate intelligence. In these cases, Morgan's Canon produces errors in the other direction. Secondly, even in the cases in which it should be most effective, the empirical evidence suggests that it may be ineffective or even increase bias.[4] Morgan's Canon should not be used to control intuitive anthropomorphism.

5. Evidentialism

[4] Also, as mentioned, intuitive anthropomorphism can influence any stage of research. A rule about model choice, like Morgan's Canon, doesn't address anthropomorphism at other stages.

The next control is proposed by Sober (2005) and Fitzpatrick (2008) as replacements for Morgan's Canon. Sober concludes his paper with the memorable line: "The only prophylactic we need is empiricism" (2005 pg. 97). Fitzpatrick frames the claim as a principle he calls 'evidentialism' (2008, pg. 242). The shared idea is that one should wait until there is sufficient evidence before adopting a hypothesis.

This, of course, is good advice. The problem is that inferences about *what is sufficient evidence* are potentially subject to intuitive anthropomorphic bias. This is, arguably, exactly what is at issue in this debate. We need guidelines that *specifically* address intuitive anthropomorphism, so evidentialism alone is not enough.

6. Identifying Errors

Many authors discuss more specific instantiations of anthropomorphic error. This suggests a third strategy. The more specific the errors we can identify, the better positioned we are to avoid them. Simply warning against 'anthropomorphism' in general is too vague to be helpful.

De Waal (1999) argues against *anthropocentric anthropomorphism*, which is attempting to explain behavior by simply imagining how you would approach the task. He emphasizes that researchers need to consider the perspective of the animal itself. Thus, there is an identifiable difference between instances of pernicious anthropocentric anthropomorphism and what he calls *animal-centric* anthropomorphism. Rivas & Burghardt (2002) describe the related *anthropomorphism by omission*, which assumes that the capacities of nonhuman animals are a subset of our own. In reality, nonhuman animals have many capacities that we do not.

Lockwood (1986) distinguishes several kinds of anthropomorphism, the most interesting of these is *explanatory anthropomorphism*. This is the fallacy of thinking we have explained a

behavior simply by giving it a (folk psychological) name. Many raise related worries about the use of folk psychological terminology, which might make it easy to unintentionally slip into intuitive anthropomorphism. Authors worry about this to different degrees. Kennedy (1992) argues for the complete replacement of folk psychological terminology, while de Waal (1999) argues for the use of neutral descriptive terminology when observing and recording data, though folk psychological can be used theoretical interpretation. Finally, Bekoff (2000) argues that general care is all that is needed.

Povinelli lists several specific errors (2012, appendix 1), for instance, the *end-point training effect*: one watching an animal perform some task can be struck by her success, and forget that the behavior required hundreds or thousands of practice trials. Another, he calls *The Erin Moriarty effect* (after a journalist who visited his lab): in tasks in which an animal has a 50/50 chance of 'success,' trials in which it succeeds *feel* more meaningful.

Whether or not these specific proposals stand up, there is a useful idea here. The problem is that there is little reason to believe that we will be able to identify a list of discrete errors that exhausts intuitive anthropomorphic bias. The effects of cognitive bias are subtle, complex, and often unpredictable. And we need controls that will generalize to experiments not yet done. So, like evidentialism, this approach is helpful, but more is needed.

7. Building New Strategies

I have argued that awareness of bias in deliberation is not an effective control itself, but it might seem hasty to argue that it has no role. Either way, though, a proper understanding of bias is essential. So given that the field has implicit anthropomorphism wrong, adopting the view that I have argued for can only help. But we need more than this. Of existing controls, Morgan's

Canon should be eliminated, while evidentialism and identifying errors can be helpful, but are not enough. So I suggest that we look to the literature on controlling social biases for new strategies.

The literature on controlling implicit biases suggests an empirical hypothesis. Imagining intelligent seeming actions performed by a being that is (relatively) unlikely to be anthropomorphised, such as a computer, an insect, or an octopus (Eddy, Gallup, & Povinelli 1993), might reduce subsequent intuitive anthropomorphism. This is analogous to imagining positive interactions with members of stereotyped groups. As an empirical claim, this needs to be tested, and testing it will not likely be easy.

In the meantime, there are proposals better supported by the evidence discussed above. Implementation intentions reduce implicit bias because they help ensure that a specific goal is implemented at the right time (Stewart & Payne 2008). For practicing comparative psychologists, this would be much more difficult because the goals will have to be much more complicated. So, in place of implementation intentions, I suggest checklists. Checklists have been helpful in many settings, including airline takeoffs, engineering, and surgery (Borchard et al. 2012), and have been suggested as a method to reduce implicit bias in judges (Seamone 2006). Checklists can ensure that a large number of complicated intentions are enacted at the right time.

This checklist might include two sets of items. One set should specify alternative hypotheses that might explain the behavior. This should include hypotheses beyond that which is favored in the field and that which is the researchers own hypothesis: Heyes (2015) argues that many well-known effects that could explain behaviors are often ignored in specific debates in comparative psychology (One could also produce new hypotheses by imagining how something unlikely to be

anthropomorphised could perform the task). Another set of items could come from a more systematic program of identifying errors, modelled on the heuristics and biases literature (e.g. Gillovich, Griffin, & Kahneman 2002). A more complete *taxonomy* of intuitive anthropomorphic errors can help in many cases, even if it is not alone sufficient. If there is an alternative hypothesis, or there is reason to believe an error has been made, one must step back and reassess the evidence.

These suggestions of the last section need to be tested, and perhaps better measures can be found, but for now, these are the options best supported by the empirical evidence we have.

8. Conclusion

One might reasonably wonder what this discussion means for the current state of comparative psychology, as I have explicitly avoided this question. This discussion does not tell us whether the field in general overestimates the intelligence of nonhuman animals. While overestimating intelligence is one effect of intuitive anthropomorphism, it is only one of them. It may not even be a dominant effect; it is easily identified, so its apparent prominence may simply be an matter of availability.

One effect that the dynamic between intuitive anthropomorphism and Morgan's Canon might have is to produce a stark divide between sophisticated 'cognitive' capacities, often couched in folk psychological terms, and overly simplistic associative explanations. This concern has been raised elsewhere (Penn 2011, Heyes 2015), but considerations raised here can (partially) explain it. Not only do folk-psychological models seem more intuitively plausible, intuitive anthropomorphism can influence the *building* of models. So, either models are

constructed using the few mathematical tools we have (like associative modeling) or under the heavy influence of folk psychology. Beyond that, I suspect there is little that can be said in general, and progress will have to be made on a case by case basis.

References

Arico, A., Fiala, B., Goldberg, R. F., & Nichols, S. (2011). The Folk Psychology of Consciousness*. Mind & Language, 26(3), 327-352.

Banks, R. R., Eberhardt, J. L., & Ross, L. (2006). Discrimination and implicit bias in a racially unequal society. *California Law Review*, 1169-1190.

Barrett, J. L., & Keil, F. C. (1996). Conceptualizing a nonnatural entity: Anthropomorphism in God concepts. Cognitive psychology, 31(3), 219-247.

Bekoff, M. (2000). Animal emotions: exploring passionate natures. *BioScience*, *50*(10), 861-870.

Bertrand, M., & Mullainathan, S. (2003). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination (No. w9873). National Bureau of Economic Research.

Bodenhausen, G. V., & Wyer, R. S. (1985). Effects of stereotypes in decision making and information-processing strategies. Journal of personality and social psychology, 48(2), 267.

Borchard, A., Schwappach, D. L., Barbir, A., & Bezzola, P. (2012). A systematic review of the effectiveness, compliance, and critical factors for implementation of safety checklists in surgery. Annals of surgery, 256(6), 925-933.

Caporael, L. R., & Heyes, C. M. (1997). Why anthropomorphize? Folk psychology and other stories. *Anthropomorphism, anecdotes, and animals*, 59-73.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. Journal of Personality and Social Psychology, 44(1), 20.

de Waal, F. B. (1991). Complementary Methods and Convergent Evidence in the Study of Primate Social Cognition1). Behaviour, 118(3), 297-320.

de Waal, F. B. (1999). Anthropomorphism and anthropodenial: consistency in our thinking about humans and other animals. Philosophical Topics, 255-280.Dyer , F.C . 2002 : The biology of the dance language . Annual Review of Entomology , 47 , 917 – 949 .

Dennett, D. C. (1989). *The intentional stance*. Cambridge, MA: MIT press.

Eddy, T. J., Gallup, G. G., & Povinelli, D. J. (1993). Attribution of cognitive states to animals: Anthropomorphism in comparative perspective. Journal of Social Issues, 49(1), 87-101.

Fiala, B., Arico, A., & Nichols, S. (2011). On the psychological origins of dualism: Dual-process cognition and the explanatory gap. Creating consilience: Integrating science and the humanities, 88-110.

Fitzpatrick, S. (2008). Doing away with Morgan's Canon. Mind & Language, 23(2), 224-246.

Gallup Jr, G. G., Marino, L., & Eddy, T. J. (1997). Anthropomorphism and the evolution of social intelligence: A comparative approach. In *Anthropomorphism, anecdotes, and animals*,

Gebhard, U., P. Nevers, and E. Billman-Mahecha. 2003. "Moralizing Trees: Anthropomorphism and Identity in Children's Relationships to Nature." In *Identity and the Natural Environment: The Psychological Significance of Nature*, S. Clayton and S. Opotow (Eds.), 91–112. Cambridge, MA: MIT Press.

Govan, C. L., & Williams, K. D. (2004). Changing the affective valence of the stimulus items influences the IAT by re-defining the category labels. Journal of Experimental Social Psychology, 40, 357 – 365.Graham 1998

Green, Alexander R., et al. "Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients." Journal of general internal medicine 22.9 (2007): 1231-1238.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. The American Journal of Psychology, 243-259.

Heyes, C. (2015). Animal mindreading: what's the problem?. Psychonomic bulletin & review, 22(2), 313-327.

Kennedy, J. S. (1992). The new anthropomorphism. Cambridge: Cambridge University Press.

Liu, J., Li, J., Feng, L., Li, L., Tian, J., & Lee, K. (2014). Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia. Cortex, 53, 60-77.

Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages How motivational interventions Can reduce (but also increase) prejudice. Psychological Science, 0956797611427918.

Lockwood, R. (1986). Anthropomorphism is not a four-letter word. In *Advances in Animal Welfare Science 1985* (pp. 185-199). Springer Netherlands.

Mitchell, S. D. 2005. Anthropomorphism and cross-species modeling. In Thinking with Animals: New Perspectives on Anthropomorphism, 100–117, ed. L. Daston and G. Mitman. New York: Columbia University Press.

Mitchell, T. L., Haw, R. M., Pfeifer, J. E., & Meissner, C. A. (2005). Racial bias in mock juror decision-making: a meta-analytic review of defendant treatment. *Law and Human Behavior*, *29*(6), 621.

Morewedge, C. K., Preston, J., & Wegner, D. M. (2007). Timescale bias in the attribution of mind. *Journal of personality and social psychology*, 93(1), 1.

Morgan, C. L. (1894). Introduction to comparative psychology. London, UK: Walter Scott Ltd.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Science*s, 109(41), 16474-16479.

Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. Journal of Experimental Social Psychology, 38(4), 384-396.

Penn, D. (2011). How folk psychology ruined comparative psychology: and how scrub jays can save it. *Animal thinking: contemporary issues in comparative cognition. MIT Press, Cambridge*, 253-266.

Povinelli, D. J. (2012). World without weight: Perspectives on an alien mind. Oxford University Press.

Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., & Guthrie, C. (2009). Does unconscious racial bias affect trial judges?. *notre dame law review*, *84*(3), 09-11.

Rasmussen, J. L., & Rajecki, D. W. (1995). Differences and similarities in humans' perceptions of the thinking and feeling of a dog and a boy. Society & Animals, 3(2), 117-137.

Rivas, J., & Burghardt, G. M (2002). Crotalomorphism: A metaphor for understanding anthropomorphism by omission. In M. Bekoff, C. Allen, & G. M. Burkhardt (Eds.), The cognitive animal: Experimental and theoretical perspectives on animal cognition (pp. 9-17). Cambridge, MA: MIT Press.

Seamone, E. R. (2006). Understanding the person beneath the robe: Practical methods for neutralizing harmful judicial biases. *Willamette L. Rev.*, 42, 1.

Shettleworth, S. (2010) Cognition, evolution, and behavior, 2nd edn. Oxford, New York.

Sober, E. (2005). Comparative psychology meets evolutionary biology. Thinking with Animals, 85-99.

Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. Personality and Social Psychology Bulletin, 34, 1332 – 1345.

Turner, R. N., & Crisp, R. J. (2010). Imagining intergroup contact reduces implicit prejudice. British Journal of Social Psychology, 49, 129 – 142.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131.

Uhlmann, E. L., & Cohen, G. L. (2007). "I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes*, 104(2), 207-223.

Wilder, H. 1996. Interpretative cognitive ethology . In M. Bekoff and D. Jamieson (eds), Readings in Animal Cognition . Cambridge MA : MIT Press .

Wynne, C. D. (2007). What are animals? Why anthropomorphism is still not a scientific approach to behavior. Comparative Cognition & Behavior Reviews, 2, 125-135.

Problems and Questions in Scientific Practice

Author: Steve Elliott, Center for Biology and Society, Arizona State University

Abstract

Philosophers increasingly study how scientists conduct actual scientific projects and the goals they pursue. But as of yet, there are few accounts of goals that can be used to identify different kinds, and specific instances, of goals pursued by scientists. I propose that there are at least four distinct kinds of goals pursued by scientists: ameliorating problems, addressing questions, satisfying values, and achieving epistemic aims. I focus on the first two kinds, and I provide tools to help conceptualize, distinguish, and identify the problems and questions pursued by scientists. This paper illustrates the use of those tools with two examples.[1]

[1] Thanks to Tom Nickles, Rick Creath, and Manfred Laubichler for comments on earlier versions of this paper.

1- Introduction

Philosophers increasingly study how contemporary scientists conduct ongoing scientific projects, often called scientific practice (Ankeny et al. 2011). Philosophers focus on the myriad theories, models, laws, hypotheses, and similar items (here grouped under the term 'scientific products') deployed in those projects (Green 2013; O'Malley et al. 2014). In doing so, they note that scientists construct, select, and deploy scientific products to achieve many different kinds of goals, from better describing phenomena, to explaining it, to addressing issues like deforestation or abnormal embryonic development (Elliott and McKaughan 2014; Love 2013). But as of yet, no one has developed a systematic account of the different kinds of goals actually pursued by scientists.

Such studies have yielded a wealth of new concepts with which to study the goal-directed aspect of much of science. Many accounts use similar terms for different concepts, such as for concepts of goals and aims (Elliott and McKaughan 2014; Potochnik 2015), and for values (Douglas 2013; Brigandt 2015), to name just a few. Older but kindred accounts conflated terms or concepts of aims with those of values (Kuhn 1977; Laudan 1984), or of question-answering with those of problem-solving (Laudan 1977; Hintikka 1981).

That florescence of concepts portends some issues. First, it may prompt philosophers to argue about, and entrench their positions against, concepts or accounts that are in fact compatible with their own. More importantly, it has made difficult at best our ability to compare results across different studies of scientific practice. Without such comparisons, such studies amount to little more than recounts of scientific research, yielding little understanding of the epistemologies of science. We need instead a framework of the myriad goals of scientific practice, combined

with a strategy with which to study and systematically compare cases of scientific practice. Here, I focus on the first issue.

To ameliorate the above issue, I propose a framework of the goals of scientific research. The framework knits together many of the concepts proposed by others. I focus on two such goals: solving problems and answering questions. Many have nearly identified the two, but I argue that we more fruitfully study scientific practice if we distinguish them.

In the next section, I sketch the overall framework of goals, while I discuss problems and questions in sections 3 and 4, respectively. In section five I provide some reasons for distinguishing problems from questions, and in section 6 I highlight two research projects that we better understand if we distinguish the questions they address from the problems they ameliorate. I close the paper by forestalling some worries about the overall sketch and about the distinction between problems and questions.

2- Goals of Science

Many accounts of the goals, aims, or ends of science privilege an ultimate goal, such as explanation (Popper 1957), problem solving (Laudan 1977), significant truth (Kitcher 2001), or understanding (Potochnik 2015). Some accounts propose multiple aims, especially focusing on addressing practical issues (Elliott and McKaughan 2014). Critics urge that there are no general aims of science, and that we should instead develop concepts of aims that are localized to researchers (Hardcastle 1999). I propose multiple goals of science focused around research teams. But my account knits together some of the insights of those who proposed ultimate aims.

I propose that contemporary researchers pursue at least four kinds of aims: to solve problems, to answer questions, to achieve epistemic goals (such as describing, explaining, or

predicting phenomena), and to satisfy values. When I focus on scientific products instead of on researchers, I use the language of functions for artifacts (Knuuttila 2011; Love 2013; Woodward 2014). Thus, scientific products function to solve problems, answer questions, etc.

Furthermore, over the course of a project, the functions of problems, questions, etc. themselves change. At the beginning of a project, they motivate researchers to act. In the middle part of a project, they constrain the behaviors of researchers. At the end of a project, researchers evaluate the results of their projects and those of other according to those problems, questions, etc.

The above sketch provides a bit of context for the sections to come. I can't here detail all elements of the sketch, so I focus on two aspects of it: problems and questions. The next two sections provide the machinery with which to conceptualize problems and questions, while the ensuing section illustrates examples in which the machinery is useful.

3- Problems

If we're to take solving problems as a general goal of many scientific projects, we need an account of problems (Nickles 1981). To apply to actual scientific practice, such an account must provide a set of tools with which to identify the problems that scientists pursue, a known but little studied issue (Nickles 1988). The set of tools should include at least the following.

First, the account must specify the kinds of things to which 'problem' refers. Second, the account must provide a general semantical structure for propositions of the form "X is a problem". Third, the account must provide conditions according to which researchers assert the proposition that 'X is a problem'. With those tools, those who study scientific practice can better identify the problems that scientists identify as parts of their projects. I sketch four tools below.

1. Problems:

   Problems are, and the term ‘problem’ refers to, states of affairs or situations in which something valued is harmed or is obstructed from flourishing.

Many who discuss problems in science adopt more restricted views of problems. They view problems merely as troubles faced by theories (Laudan 1977; Hattiangadi 1978; Nickles 1981). That view prompts most to treat problems as questions, a positon I discuss in a later section. My account is more general, noting that scientists often motivate and evaluate their projects by more worldly issues, such as droughts, birth defects, and extinctions.

2. Proposition that “X is a problem”:

   The proposition is an abstract object that includes: a set of propositions that describe a situation, an evaluative proposition that disvalues the situation, an imperative proposition to ameliorate the situation, and a set of propositions that describe constraints on the amelioration.

The above tool is due largely to Thomas Nickles (1981). My version is more general than his in several ways. First, I countenance more kinds of situations as problems than does Nickles. Second, for Nickles, all scientific problems include at least obstructed goals in relation to theories; while my account countenances harms or obstructions to explicit theoretical goals, but also to anything valued. Third, my account explicitly includes propositions to describe states of

affairs and an evaluative proposition, while Nickles's account at best lumps those propositions as kinds of constraints on solutions.

3. Problem Statement:

   A problem statement is a sentence that describes a problem. Put differently, a problem statement expresses the proposition "X is a problem" and specifies the X.

A problem statement uses the concept of a problem when it implies all of the propositions that constitute that concept. For instance, the sentence "Deforestation is a problem in Montana" implies, among other things, that deforestation in Montana is dis valuable and that it should be reduced or halted.

Nickles aimed to describe problems as things that could persist across projects and researchers, that could evolve over time, and that themselves could be studied. While he noted that agent-focused tools might complement his semantic account and be useful to study scientific practice, he provided none. The following is one such complementary tool. It provides a set of conditions in which agents can assert a problem statement.

4. Agents (straightforwardly) assert that "X is a problem" only when:
   1. They're conscious or aware of the state of affairs X
   2. They disvalue X
   3. They imply an imperative to ameliorate X
   4. They believe that effort is needed to further specify or to ameliorate X
   5. They believe that a possible strategy of action could ameliorate X

6. They believe that, when pursuing problems, it's appropriate to use the language of ameliorable (ability), ameliorating (process), and ameliorated (result).

The above tool is largely due to Gene Agre (1982). According to it, agents assert that some situation is a problem based on their background knowledge, values, and beliefs, all of which can differ across agents, and the latter two of which can be disputed. Furthermore, it provides a starting point around which those who study scientific practice can develop tools, such as surveys and content analyses, to collect data about the problems invoked and pursued by researchers.

In a further respect, the above tools are more general than nearly all earlier accounts of problems. Earlier accounts describe researchers as solving problems. The above tools instead describe them as ameliorating problems. More than language of "solving", language of "ameliorating" better fits the model of satisficing rationality explicitly invoked by most previous accounts of problems.

4- Questions

If we're to take answering questions as a general goal of many scientific projects, we need an account of questions. There has been substantially more research into questions than into problems, yielding a wide array of topics and accounts (Cross and Roelofsen 2016). Given that array, much less machinery needs to be developed to identify questions pursued in scientific practice. For semantics of questions, I review a standard account familiar to philosophers of

science (Belnap and Steel 1976), while for conditions under which agents poses questions, I specify a new account.

Questions are abstract objects, like propositions, posed by agents. For their most basic semantic structure, elementary questions include a set of possible alternatives and propositions that indicates how many of the distinct alternatives the agent seeks (Belnap and Steel 1976). An interrogative statement is a sentence that expresses a question, just as a problem statement expresses a problem proposition.

For a question to possibly have an answer, it presupposes a proposition that describes a state of affairs in which the agent asking the question lacks information. Agents pose questions under at least the following conditions.

5. Agents (straightforwardly) pose a question only when:
    1. They're aware of their epistemic state of lacking information
    2. They disvalue that state
    3. They imply an imperative to ameliorate that state
    4. They believe that effort is needs to formulate the question or to provide the information to address it
    5. They believe that possible information could ameliorate their epistemic state, and that they could identify that information if presented it (Hintikka 1981).

6. They believe that, when pursuing questions, it's appropriate to use the language of addressable (ability), addressing (process), and addressed (result).[2]

There are many parallels between the above accounts of problems and questions, and I model the conditions for posing questions on Agre's conditions for asserting problem propositions. Such similarities between the accounts partly explains why so many philosophers have given the appearance of equating the two. But for studying science, we most fruitfully treat problems states of affairs and questions as abstract objects.[3]

5- Distinguishing Problems from Questions

Among those who study the role of problems in science, many seemingly identify problems with (sets of) unanswered questions and the practice of problem solving with that of questions answering (Laudan 1977, Hintikka 1981, Goldman 1986, Love 2008). The tools in the previous sections enable us to distinguish problems from questions, and to distinguish the practice of ameliorating problems from that of addressing questions.

[2] Language of 'addressing questions' replaces that of 'answering questions' to better fit with the satisficing model of rationality.

[3] Insofar as one ontologizes states of affairs as themselves abstract objects, then my argument is instead that problems have logical or abstract structure that is distinct from that of questions. Thus, the set of problems doesn't overlap with the set of questions.

There are several reasons why it is fruitful to distinguish questions from problems when we study research projects. First, the tools described above enable us to charitably interpret previous accounts of problems such their insights are still highly relevant to the study of scientific practice. Given those tools, questions presuppose a kind of problem proposition, and the practice of addressing questions is a subkind of the practice of ameliorating problems. Given a question, we might establish some translation rule to move between the question and a specific one of its presuppositions and back again.

For instance, if I ask "Why do leafblowers make as much noise as they do?", I presuppose that I lack that information, a situation I disvalue. If we treat lacking information as a situation in which something valued (here knowledge, information, or understanding) is obstructed, then we can class the situation as an (intellectualist) problem. Given appropriately developed translation procedures, we could infer "I lack information about what causes leafblowers to make as much as they do" from the above question. And given a proposition that describes an agent's epistemic state of lacking information, we could infer a question from that proposition. Briefly put, a question presupposes an intellectualist problem proposition, and with the right translation procedures, we could move back and forth between the two.

But we cannot develop such translation procedures for problems that aren't epistemic states of lacking information. Such problems engender many questions. If an agent asserts that leafblowers are noisy, we can't infer that the agent doesn't know why they are noisy. She may know perfectly well, and she may be developing a noise damper. Rather, the problem engenders many questions: How noisy are they? For whom are they noisy? Where are they noisiest? When are they nosiest? What is their range of noise? How can we dampen the noise? Why do people

use leaf blowers? Etc. For every answered question, people gain information with which they can ameliorate the problem.

Older accounts of problems and questions captured an aspect of many problems pursued in scientific practice. In such cases, researchers disvalue their situations of lacking knowledge, situations that can be expressed either as problem propositions or as presuppositions to questions. The distinctions I offer enable us to maintain most of the usefulness of those accounts, such as rough distinctions between conceptual and empirical problems (Laudan 1977), the abstract structure of problem propositions (Nickles 1981), and an interrogative account of scientific discovery (Hintikka 1981).

There's a second general reason for why it's fruitful to distinguish problems from questions as I do above. The distinction enables us to conceptualize, beyond the epistemic-state or intellectualist problems presupposed by questions, a larger variety of problems that researchers pursue and that give significance to their research projects. Such worldly problems are aspects of projects that comprise a vast expanse of scientific practice, including the study of fault lines near cities, extinctions, the physical decay of art, and disease, to name just a few. Often ignored by those who study science, such research has the potential to reveal not only how researchers design and conduct research, but also how they make trade-offs between competing values (Elliott and McKaughan 2014).

Third, the distinction I propose enables the creation of tools to identify the problems and questions pursued by researchers. Insofar as we study science empirically, we need empirical tools to gather data about the aspects of research projects and how those aspects evolve over time.

For instance, the two sets of conditions, one for asserting that something is a problem and the other for posing a question, provide foundations from which to create questionnaires or surveys of researchers about the problems and questions they pursue, foundations that are themselves revisable in light of disconfirming evidence. Similarly, the abstract structures of problem propositions and of questions provide foundations from which to create content analyses of documents such as research papers, grant applications, lab notebooks, etc.

Fourth, the distinction between questions and problems enables the study of how those problems and questions function differently over the course of a given research project, and how they influence each other. Both can motivate scientists to design and conduct a project. But while questions focus research activity and function as criteria by which to select methods, problems often function as external justification for a course of action. Further, problems often engender or raise many questions, while the information used to address questions often ameliorates many distinct problems.

A team might be motivated to conduct their project because of one problem, but may invoke a different problem when applying for funding. It may ignore both of those problems and invoke still a different problem when reporting its results and convincing others to use its scientific products. Problems are not, however, mere rhetorical devices, as much of science evolves by using the same results or products to ameliorate different problems, and many products are evaluated by how well they ameliorate problems, and by how many problems they ameliorate.

Finally, the distinction between questions and answers enables more refined studies not only into how science is or should be evaluated and conducted, but also into how it is or should be designed. In that sense, the distinction is foundational in a general study of the conceptual

foundations of research design, a field scarcely touched by philosophers of science, and one ripe for further study.

6- Examples

I describe two examples that highlight the above tools and the distinction between questions and problems. The first example straightforwardly fits my account, while the latter, while prima fascia challenging to my account and more amenable to older accounts, also fits it.

Example 1

In Death Valley in California and Nevada, there are many species of pupfish isolated from each other in streams and water holes. One species lives only in Devils Hole, a seemingly bottomless, hot, and geothermal hole only a few meters in width and breadth that sustains little life. Devils Hole pupfish are distinct from pupfish in sister species in that they lack pectoral fins. The pupfish eat mostly the small amounts of algae that grow in Devils Hole, and when shade ceaselessly occludes the hole for two months every year, algae can't grow and the pupfish population crashes. While the population rarely numbers more than a few hundred, during the shady season it has been recorded at barely a few dozen.

The Devils Hole species is one of the most endangered on the planet. The US federal government has explicitly valued and protected not only the species, but also the hole and the water in which it lives (Cappaert v. United States 1976). The US Fish and Wildlife Service, among other federal organizations, pursues several efforts to conserve the species.

In the framework of problems proposed in earlier sections, the population of Devils Hole pupfish is the thing valued, and its constant threat of extinction provides a problem at least to the

Fish and Wildlife Service tasked with conserving it. That problem has engendered many questions about the pupfish.

A recent team to study the pupfish was led by Christopher Martin, a specialist in the speciation of small fishes at the University of North Carolina, Chapel Hill. Motivated by the conservation problem facing the Fish and Wildlife Service, the team primarily asked: How long had the Devils Hole species lived in Devils Hole? But it also noted a cluster of related questions (Martin et al. 2016, 3). Did pupfish colonize the hole just once? If so, how did their populations avoid inbreeding depression and extinction?

To answer those questions, the team sampled DNA from preserved Devils Hole pupfish and compared it to DNA data from nearby sister species. They used that information to build several scientific products, including a metric of genetic diversity in and between species, a dated phylogenetic tree of the species, a DNA mutation rate, and a time-range in which the Devils Hole pupfish diverged from its sister species.

Given those tools, the team answered it primary question by inferring that the current species of Devils Hole pupfish colonized Devils Hole within the last three hundred years, a surprisingly recent event. And given geological record of inundations in Death Valley, the team concluded that the current species of Devils Hole pupfish may be just the most recent species to colonize the hole. They also found evidence of gene flow between the Devils Hole population and sister species, though kilometers of desert often separated those populations. Such gene flow could stave off inbreeding depression, though mechanisms are needed to explain how genetic material somehow traversed kilometers.

With its scientific products and answers to their questions, the team recommended some strategies to conserve the Devils Hole pupfish, or to ameliorate the problem facing the Fish and

Wildlife Service. The team suggested that the current species may be only the most recent in a series of species that have colonized Devils Hole, evolved, and gone extinct. To conserve pupfish in Devils Hole, if not the current species, fish and wildlife managers should preserve the possibility for genetic information, and perhaps new organisms, to flow to and from Devils Hole. Whether or not anyone employs those suggestions to ameliorate the problem of potential extinction of the pupfish, or of life in Devils Hole, remains to be seen.

Example 2

Not all research projects fit my account as nicely as the previous case. When many scientists describe their projects, they often don't identify questions or worldly problems as motivation or justification for their projects. They often invoke intellectualist problems, such as a lack of understanding of a phenomenon, or trouble for a theory. Such cases are those traditionally studied by those who study problem solving in science. But as my account conserves and clarifies previous accounts, it can still handle such cases.

In the late 1960s, Eric Davidson partnered with Roy Britten to develop a project about gene regulation. They noted evidence that genes yielded products that regulated how other genes made products, and ultimately how cells differentiate. Davidson and Britten valued knowledge of how genes control cell differentiation. And the problem, as they came to state, was that researchers knew little about those mechanisms in which genes regulate each other, and they had little theory in which to describe such mechanisms or figure them out (Britten and Davidson 1969).

To address that problem, Britten and Davidson proposed a set of scientific products, which included theoretical concepts and a general model of a gene battery, according to which

gene products within a cell interacted with other genes in the cell to differentiate the cell into a given type. A few years later, they moved to Caltech to collaborate regularly. They pursued major grants together, and over time, their project and its motivating problems evolved. They helped to establish many instances in which genes regulated other genes in sea urchins.

But by the late 1990s, the theory had evolved enough that Davidson and an army of colleagues could pose a new problem. No one had provided a relatively complete example of a gene regulatory network, as the batteries had been renamed, according to which researchers could manipulate gene regulation and precisely predict the effects on a major developmental process. Davidson was aware of that situation, disvalued it as a challenge to his and Britten's theory, and exercised a considerable amount of effort to ameliorate it via a complex strategy (Davidson et al. 2002).

To address that problem, Davidson's team focused on the specification of endomesoderm cells in early sea urchin embryos, some of which develop into the distinctive juvenile skeletons found in many sea urchin larvae. The team systematically perturbed the expression of dozens of genes. Ultimately, they constructed a model of all the genes and their regulatory connections required to turn early sea urchin cells into endomesoderm cells, and ultimately juvenile skeletons. Given a model that ameliorated, but didn't completely solve, the above problem, Davidson's research evolved as he continued to refine motivating problems, from needing more detail on the endomesoderm network (Peter and Davidson 2010) to not knowing how it had evolved (Hinman et al. 2007).

7- Conclusion

In closing, I forestall some worries about the above sketch of problems and questions. First, while it characterizes problems and questions, is says little about solutions and answers, or about how researchers go about finding them. So the above sketches are incomplete. That point is right, and while a more thorough study of those topics is beyond the scope of this paper, it is ripe for future research.

Second, my examples draw evidence only from the published reports of scientists, reports that are known not to capture how their authors actually reasoned during the life of their research projects. As such, the above examples may systematically mislead readers about the problems and questions actually pursued in the projects, which the authors rationally reconstruct in their reports. To an extent, that point is also right. But it misses at least two larger issues.

First, within any given research project, the functions of problems and questions change over the life of the project. In early stages, they motivate researchers to act, in later stages they constrain the behaviors that researchers perform, and in still later stages they justify the project, its results, and its products to other scientists. Here, I focus my examples only on the late stage in which researchers publish their results. Fuller case histories, however, would identify the problems and questions, and how they changed, throughout the life of the projects.

Second, while scientists reconstruct their projects in their research reports, that practice is a worthwhile object of study. Those who study them should be mindful that such reconstructions sacrifice historical accuracy and have rhetorical functions. But those reconstructions, and the practice of making them, provide a window through which those who study science can piece together the rationales for projects, results, and scientific products that scientists find convincing. Problems and questions are often distinct aspects of those rationales.

References:

Agre, Gene P. 1982. "The Concept of Problem." *Educational Studies* 13: 121–42.

Ankeny, Rachel, Hasok Chang, Marcel Boumans, and Mieke Boon. 2011. "Introduction: Philosophy of Science in Practice." *European Journal for Philosophy of Science* 1: 303–7.

Belnap, Nuel D., and Thomas B. Steel. 1976. *The Logic of Questions and Answers*. New Haven, Conn.: Yale University Press.

Brigandt, Ingo. 2015. "Social Values Influence the Adequacy Conditions of Scientific Theories." *Canadian Journal of Philosophy* 45: 326–56.

Britten, Roy J., and Eric H. Davidson. 1969. "Gene Regulation for Higher Cells: A Theory." *Science* 165: 349–57.

Cappaert v. United States, 426 US 128 (Supreme Court 1976).

Cross, Charles, and Floris Roelofsen. 2016. "Questions." Edited by Edward N. Zalta. *The Stanford Encyclopedia of Philosophy* http://plato.stanford.edu/entries/questions/ (Accessed February 26, 2016).

Davidson, Eric. H., et al. 2002. "A Genomic Regulatory Network for Development." *Science* 295, no. 5560: 1669–78.

Douglas, Heather. 2013. "The Value of Cognitive Values." *Philosophy of Science* 80: 796–806.

Elliott, Kevin C., and Daniel J. McKaughan. 2014. "Nonepistemic Values and the Multiple Goals of Science." *Philosophy of Science* 81: 1–21.

Goldman, Alvin. 1986. *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Green, Sara. 2013. “When One Model Is Not Enough: Combining Epistemic Tools in Systems Biology.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 44: 170–80.

Hardcastle, Gary. 1999. “Are There Scientific Goals?” *Studies in History and Philosophy of Biological and Biomedical Sciences* 30: 297–311.

Hattiangadi, J.N. 1978. “The Structure of Problems, (Part I).” *Philosophy of the Social Sciences* 8: 345–65.

Hinman, Veronica F., Albert Nguyen, and Eric H. Davidson. 2007. “Caught in the Evolutionary Act: Precise Cis-Regulatory Basis of Difference in the Organization of Gene Networks of Sea Stars and Sea Urchins.” *Developmental Biology* 312: 584–95.

Hintikka, Jaakko. “On the Logic of an Interrogative Model of Scientific Inquiry.” 1981. *Synthese* 47: 69–83.

Kitcher, Philip. 2001. *Science, Truth, and Democracy*. New York: Oxford University Press.

Knuuttila, Tarja. 2011. “Modelling and Representing: An Artefactual Approach to Model-Based Representation.” *Studies in History and Philosophy of Science Part A* 42: 262–71.

Kuhn, Thomas. 1977. *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago: University of Chicago Press.

Laudan, Larry. 1977. *Progress and Its Problems*. Berkeley: University of California Press.

———. 1984. *Science and Values*. Berkeley: University of California Press.

Love, Alan C. 2008. “Explaining Evolutionary Innovations and Novelties: Criteria of Explanatory Adequacy and Epistemological Prerequisites.” *Philosophy of Science* 75: 874–86.

———. 2013. “Theory Is as Theory Does: Scientific Practice and Theory Structure in Biology.” *Biological Theory* 7: 325–37.

Martin, Christopher H., Jacob E. Crawford, Bruce J. Turner, and Lee H. Simons. 2016. “Diabolical Survival in Death Valley: Recent Pupfish Colonization, Gene Flow and Genetic Assimilation in the Smallest Species Range on Earth.” *Proceedings of the Royal Society B: Biological Sciences* 283: 20152334.

Nickles, Thomas. 1981. “What Is a Problem That We May Solve It?” *Synthese* 47: 85–118.

———. 1988. “Questioning and Problems in Philosophy of Science: Problem- Solving Versus Directly Truth-Seeking Epistemologies.” In *Questions and Questioning*, edited by Michel Meyer, 43–67. New York: Walter de Gruyter.

O’Malley, Maureen A., Ingo Brigandt, Alan C. Love, et al. 2014. “Multilevel Research Strategies and Biological Systems.” *Philosophy of Science* 81: 811–28.

Peter, Isabelle S., and Eric H. Davidson. 2010. “The Endoderm Gene Regulatory Network in Sea Urchin Embryos up to Mid-Blastula Stage.” *Developmental Biology* 340: 188–99.

Popper, Karl R. 1957. “The Aim of Science.” *Ratio* 1: 24–35.

Potochnik, Angela. 2015. “The Diverse Aims of Science.” *Studies in History and Philosophy of Science Part A* 53: 71–80.

Woodward, James. 2014. “A Functional Account of Causation.” *Philosophy of Science* 81: 691–713.

Robust Realism for the Life Sciences

Contributed paper, PSA2016

Markus I. Eronen

Center for Logic and Analytic Philosophy, KU Leuven

Markus.Eronen@kuleuven.be

Abstract

According to entity realism, we are warranted in believing that some entities studied by science are real, but not that scientific theories are true. In discussions of scientific realism, entity realism is usually quickly dismissed due to serious objections that appear to make it untenable. In this paper, I formulate a new robustness-based version of entity realism, and show that this version has resources to answer the classic objections raised against the original version. I also show that, in contrast to the currently popular (ontic) structural realism, robustness-based entity realism provides a plausible account of realism for the life sciences.

Word count: 4960

1. Introduction

The core idea of standard scientific realism is that we ought to believe that the best scientific theories are (approximately) true. Antirealists deny this. Entity realists accept the antirealist tenet that we are not required to believe in the truth of scientific theories, but also defend limited realism, as they argue that we are warranted in believing that at least some entities that appear in scientific explanations are real. Thus, entity realism (ER) appears to lead to an appealing middle path between standard scientific realism and antirealism.

However, in discussions of scientific realism, ER is usually quickly dismissed (see, e.g., Devitt 2005; Ladyman and Ross 2007; Psillos 1999). This is probably due to two main factors: the ambiguity and inconclusiveness of the arguments for ER, and several serious counterarguments that have been raised against it. In this paper, I will formulate a new robustness-based version of ER, and show that this version has resources to answer all the classic objections raised against original ER. I will also show that, in contrast to the currently popular (ontic) structural realism, robustness-based ER provides a plausible account of realism for the life sciences.

In the next section, I will briefly go through ER and its main problems. In Section 3, I will present the robustness argument for ER, and in Section 4, I will argue that robustness-based ER has resources to answer all the main counterarguments raised against original version. In Section 5, I briefly consider the relationship between robustness-based ER and (ontic) structural realism.

2. Entity Realism

The most important accounts of ER are in Nancy Cartwright's (1983) *How the Laws of Physics Lie* and Ian Hacking's (1983) *Representing and Intervening*. I will mainly focus on Cartwright's version of ER here, as it is more compact, and for the purposes of this paper, the differences between the accounts are inessential.

Cartwright's (1983) starting point is inference to the best explanation (IBE), which is one of the classic strategies to argue for scientific realism. The core of the IBE argument is that if theories or laws are extremely successful at explaining and predicting phenomena, we can infer that they are also (approximately) true. This strategy has been forcefully criticized by a broad range of authors, most prominently Bas van Fraassen (1980). Cartwright mostly agrees with the critics, but argues that there is one exception: causal explanation. In the context of causal explanation, IBE is justified: "To the extent that we find the causal explanation acceptable, we must believe in the causes described" (Cartwright 1983, 5). In other words, she states that "to accept the explanation is to admit the cause" (ibid., 99). To illustrate, she gives the following example (ibid., 91). The lemon tree in her garden is sick, and the leaves are falling off. She comes up with an explanation: Water has accumulated at the bottom of the pot, which has made the tree sick. According to Cartwright, accepting this explanation as correct requires believing that the cause (water at the bottom of the pot) is real.

Importantly, Cartwright (1983, 75-76) also argues that accepting a causal explanation and the reality of the cause does not require accepting any theory or law as true. According to her, there is theoretical "redundancy" in science in the sense that the same causal process can be embedded into different theoretical frameworks, and consequently the reality of the causal process does not imply the truth of any theory. However, causal explanations themselves are "non-redundant", as only one causal story for a given phenomenon can be accepted as satisfactory.

Cartwright (1983) and Hacking (1983) also appeal to scientific practice and experimentation for support. Hacking famously argues that the best evidence for the reality of electrons is that we can use them to create and study other phenomena – "if you can spray them, they are real" (Hacking 1983, 23). Cartwright claims that experimentation can give causal explanations a degree of objectivity that is impossible to reach for laws and theories, referring for example to a laser-making company that runs numerous test lasers to death each year to make sure that the lasers produced have exactly the effects that they are supposed to have (Cartwright 1983, 3).

Also transcendental or indispensability arguments for ER can be extracted from the accounts of Cartwright and Hacking (Morrison 1990; Miller forthcoming). The idea is that successfully manipulating and controlling nature with scientific means *requires* accepting the reality of the entities involved. As Hacking puts it, scientists cannot help being realists about (experimental) theoretical entities (Hacking 1983, 262).

Finally, an important consideration in favor of ER is that entities seem to be more stable and resistant to scientific revolutions than theories and laws. For example, the electron entered the ontology of physics in the late 19th century, and has remained there since, although theories in physics have gone through such dramatic changes as the discovery of quantum mechanics and the relativity theory (Hacking 1982). Thus, ER appears to be less susceptible to the pessimistic induction argument than standard scientific realism.

ER is an attractive position, as it seems to amount to "defensible middle ground" between antirealism and standard scientific realism (Clarke 2001). However, ER has few proponents nowadays. This is probably due to two main reasons: the elusive nature of the arguments presented in favor of it, and serious counterarguments that can be raised against it. The elusiveness of the arguments for ER should be clear from the above summary. There is no compelling master argument, just several interconnected strands of reasoning presented in its support (see also Clarke 2001 and Miller forthcoming). I will now proceed to discuss the main counterarguments.

First of all, the notion that accepting a causal explanation as correct implies accepting the reality of the cause can be questioned. An antirealist along the lines of van Fraassen (1980) could argue that when a scientist accepts a causal explanation, she does not have to accept the reality of the cause, but merely that there is a causal story that is empirically adequate (Hitchcock 1992). Another strategy for an antirealist would be to argue that if Cartwright is correct that statements of the kind "P causally explains Q" imply the reality of P, then we are not warranted in believing that such statements are true (ibid.). There seems to be no

compelling reason why accepting the scientific practices of causal explanation would require accepting the reality of the causes (see Clarke 2001 and Hitchcock 1992 for more).

A second problem for ER can be formulated as a dilemma: Either ER is too minimal to be interesting, or it leads to a form of standard scientific realism (Morrision 1990; Psillos 1999, 249). More specifically, if ER amounts to just being warranted in believing that some entity X exist, and nothing more than that, it is questionable whether it constitutes a substantial and interesting form of scientific realism. At least, a realist also has to accept the reality of some key properties of the entity, for example, that the electron has a negative electric charge. However, the only scientifically acceptable way to attribute properties to entities is to do this on the basis of the most state-of-the-art scientific theories. This, in turn, seems to require believing that those theories are to some extent true (Chakravartty 2007, 30; Psillos 1999, 248-249). But accepting that the best scientific theories are to some extent true is not very far from accepting standard scientific realism. Thus, if entity realism is extended to cover also the properties of entities, it may not be so different from standard scientific realism after all.

A third and related problem is that, in spite of appearance, ER may not fare any better than the alternatives against the argument from pessimistic induction (Morrison 1990; Chakravartty 2007, 32). Although it may be true that, for example, the electron as such has withstood several scientific revolutions, views about its properties have considerably changed. Thomson and Rutherford believed very different things about electrons than scientists do today (ibid.). For example, they did not believe that the electron is a fermion

or that it exhibits wave-particle duality, which are nowadays seen as fundamental properties of electrons. If the views about the nature and properties of the electron have dramatically changed since it was discovered, the same presumably holds for entities as well, and the continuity that ER provides is only illusory.

In sum, the main problems of ER can be paraphrased as follows. (1) The success of causal explanation does not warrant inferring the reality of the cause. (2) Realism about entities only (without their properties) is too weak to be interesting, but the properties of entities are attributed to them by the best scientific theories, so accepting their reality involves accepting that the best scientific theories are to some extent true, leading to a form of standard scientific realism. (3) The properties of entities cannot in general be expected to survive scientific revolutions, and thus any interesting form of ER fails to avoid the pessimistic induction argument. In the next section, I will formulate a new robustness-based version of ER, and in Section 4 I will show that it successfully tackles these problems.

3. Robustness

As we saw above, Cartwright and Hacking present several interconnected arguments in support of ER: causal considerations, arguments from experimentation, indispensability arguments, and so on (see Clarke 2001 and Miller forthcoming for more). These arguments clearly have not convinced the philosophical community, as ER remains an unpopular position. However, there is a further argument for the reality of scientific entities, which

can be extracted from different debates in philosophy of science, and is far more promising for defending entity realism. This is the argument from robustness: Roughly, if there are several independent way of measuring, detecting or deriving something, then we have good reasons to believe that that thing is real.

Robustness in its different manifestations has been extensively discussed in recent years (Eronen 2015; Kuorikoski, Lehtinen, and Marchionni 2010; Kuorikoski & Marchionni forthcoming; Raerinne 2013; Schupbach forthcoming; Soler, Trizio, Nickles, and Wimsatt 2012; Woodward 2006). It is also often briefly referred to in discussions of scientific realism, including those of Cartwright and Hacking (Cartwright 1983, 84; Hacking 1981; 1983, 201; see also Chakravartty 2007, 65-66). However, it has not been developed to a full argument for scientific realism, with the exception of the work of William Wimsatt (1981, 1994, 2007), which I will take as the starting point here.

Wimsatt (1994) explicitly argues that we should adopt robustness as a criterion for what is real, and that this leads to scientific realism that is less metaphysical and more local than the standard forms. The idea is that if there are many ways of measuring, detecting, producing or deriving something, and those ways are sufficiently independent, then it is very unlikely that all of them turn out to be mistaken or erroneous. Thus, things that are robust in this sense are very likely to be real. For example, electrons can be measured, detected and produced with many different techniques and setups relying on different theoretical assumptions, and they can be derived from various models and theories. Consequently, they are robust and extremely likely to be real.

Wimsatt's rough idea is prima facie plausible, but is in several respects unsatisfactory or at least incomplete, and needs to be further refined (see also Eronen 2015). First of all, in order to avoid the implication that robustness itself makes things real (leading to some kind of constructivism), we should not see it as *criterion* for what is real in any strong sense, but rather as a source of justification or warrant for ontological commitments. This can be formulated as follows: *Robustness confers justification for believing that X is real, and the degree of this justification corresponds to the degree that we have robust evidence for X*. Furthermore, as robustness depends on currently available methods of measuring, detecting or deriving something, it is clearly a feature that is relative to a certain scientific community at a certain time. This needs to be incorporated into any definition of robustness. For similar reasons, we should take into account that robustness is a matter of degree: for example, we have more robust evidence for electrons or DNA molecules than we have for the Higgs boson. With these considerations in mind, we can give the following working definition of robustness (based on Eronen 2015):

> (Robustness) The relevant scientific community at a certain time has robust evidence for X insofar as X is detectable, measurable, derivable, producible or explanatory in a variety of independent ways.

The notion "explanatory" has been included in the definition for the reason that it is very plausible that things that appear in many independent explanatory generalizations or models are more robust[1] than things that do not (see also Eronen 2015). For example,

1 Strictly speaking, it would be more accurate to always write "robust evidence for X" instead of "x is robust", but for the sake of readability, I also use the latter kinds of

electrons are very robust partly because they appear in a broad range of distinct models and explanatory generalizations in physics, whereas D-branes only appear in certain string theory models, and are in this respect less robust. It is also important to note that none of the dimensions mentioned is by itself necessary for robustness: The moon, for example, is an extremely robust entity, although there is no clear sense in which we can produce it, and properties or phenomena can be highly robust even though there are no accepted explanatory generalizations or models involving them (e.g., gamma-ray bursts).

The notion of independence is crucial for robustness: If different ways of measuring something are not independent from each other, but are based on the same assumptions and methods, then they all lose their value if those assumptions and methods turn out to be false or mistaken, and the robustness that they confer is only illusory. One problem for robustness-based realism is that spelling out the nature of this independence is far from trivial, and if it is unsuccessful, the plausibility of the whole account can be questioned (Hudson 2014; Stegenga 2009). However, in recent years much progress has been made in defining the right kind of evidential independence. What is certainly not required is statistical independence, as two distinct ways of measuring the same thing will be often correlated, and this should not prevent them from contributing to robustness (Schupbach forthcoming). The idea is rather that two ways are appropriately independent if their characteristic errors and biases are independent from each other (Kuorikoski and Marchionni forthcoming). For example, cloud chamber experiments to detect electrons are based on different causal processes and theoretical assumptions than cathode ray tube

---

expressions here.

experiments, and thus they cannot involve the same biases or systematic errors. Ways of detection that are independent in this sense make it more likely that the entity or phenomenon is real, and thus contribute to robustness (see Kuorikoski & Marchionni (forthcoming) for more, and Schupbach (forthcoming) for an alternative proposal).

For the purposes of this paper, let us assume that the account of robustness presented here is roughly correct, so that we can examine what consequences it has for the issue of entity realism. In fact, the consequences are rather straightforward. First, it is clear that many entities in science are detectable, measurable, derivable, producible, or explanatory in a variety of independent ways, and thus we have a high degree of robust evidence for them. From this it follows that we have a high degree of justification in believing that many entities in science are real, which amounts to a form of ER. Thus, if we understand the role of robustness as I have proposed here, it directly leads to ER.[2] In the next section, I will clarify this robustness-based entity realism (from now on, RER) further, and show how the criticism raised against original ER fails to undermine it.

4. Neurons and Robustness-based Entity Realism

2 Note that the account defended here does not imply that robustness is necessary for being justified in believing something to be real: There may be also other sources of justification for ontological commitments. One important consequence of this is that cases where we are apparently warranted in believing in the reality of something that is not robust are not counterarguments to this account.

The main example that I will use here to elaborate on RER is the neuron. This is a suitable case, as the neuron is a "theoretical" entity in the sense that it is not directly observable, and in the 19th century the existence of neurons was still just a hypothesis, but nowadays there is overwhelmingly robust evidence for their reality. Here it suffices to mention just some examples of the variety of independent evidence for neurons: they can be observed with a broad range of staining techniques; they can be seen with light microscopes and imaged with electron microscopes; their activity can be recorded with various single-cell and multi-unit recording setups; they can even be produced with the help of stem cells; they play an important role in explanatory models and generalizations concerning animal and human behavior, and so on. Even if broad categories of these sources of evidence would turn out to be mistaken, plenty of other independent sources would still remain, and we would still have highly robust evidence for the neuron.

With this example in mind, let us go through the four objections to ER outlined in Section 2. The first problem was that accepting a causal explanation does not require accepting the reality of the cause, *contra* Cartwright (1983). However, in contrast to the original ER, RER does not appeal to any special features of causal explanation. In the picture I have sketched above, the fact that an entity or a property appears in a causal generalization can contribute to its robustness (as "explanatory" is one of the dimensions in the definition of robustness), but just as one possible factor among many others. A robustness-realist can accept that causal explanations are as fallible as any other explanations in science.

However, even though RER evades this particular problem, an analogous anti-realist objection can be formulated for robustness. A constructive empiricist in the vein of van Fraassen could insist that there is no compelling reason why anyone would be required (as opposed to permitted) to believe in entities for which we have robust evidence (see also van Fraassen 1985, 297-300). This may be strictly speaking true, but in the case of entities like neurons, such suspension of belief comes close to outright skepticism. It could be argued that someone who has access to all the robust evidence for neurons is just as justified in believing in the reality of neurons as in the reality of the table in front of her (see also Hacking 1981). However, as I have pointed out above, RER does not require accepting any particular theories as true, or accepting IBE as valid, so many aspects of the constructive empiricism of van Fraassen (1980) are in fact compatible with RER.

The second problem for ER was the that realism about entities only is too weak to be interesting, but realism about the properties of entities seems to require accepting that the best scientific theories are to some extent true, leading to a form of standard scientific realism. First of all, RER can and should be extended to properties as well.[3] The electrical conductivity of iron is detectable, measurable, derivable, producible and explanatory in a broad range of independent ways, and is thus an extremely robust property. Having a voltage gradient is an extremely robust property of the neuron, transmitting action

3 Original ER was also never intended to apply only to the entities themselves. For example, Cartwright explicitly states that we are warranted in believing in the reality of many “theoretical entities and *theoretical properties*” (Cartwright 1983, 8, emphasis added).

potentials is an extremely robust property of the axon, and so on. However, extending RER to properties has only minimal implications for the truth of theories. Many ways of detecting or measuring the properties of entities such as neurons do not depend on any theory. For example, Golgi's staining method for observing neurons was developed over 100 years and is still in use, but there is no accepted theory that would explain how it actually works (Guillery 2005, 1290). Furthermore, the requirement of independence guarantees that highly robust evidence for an entity or property does not rely on just one theory, but on many distinct models or theories. Any one of these models or theories may turn out to be false, and the property would still remain robust. For example, even if the Hodgkin-Huxley model for the action potential would turn out to be fundamentally incorrect, plenty of other sources of independent evidence for the action potential would still remain.

Thus, a high degree of robustness and the consequent justification in the reality of a property does not imply belief in the truth of any theory. At best, the robustness realist may be required to believe that there are some true elements among the various theories and models involved, but she can still remain entirely agnostic about the truth of scientific theories in general, and deny the validity of the IBE argument (i.e., deny that we can infer the truth of scientific theories from their explanatory success).[4]

4 One might object that robustness reasoning also involves a form of IBE: The best explanation for the robust evidence for X is that X is real, so we are justified in believing that X is real (cf. Hudson 2014). However, it is possible to accept a certain kind of IBE as valid, without accepting that IBE generally and universally works (Clarke 2001). A

The third problem was that although some entities such as the electron have withstood several scientific revolutions, many of their properties have been eliminated, and thus ER fails to evade the pessimistic induction argument. However, this issue can be reformulated and examined in new light once we understand that we can also have varying degrees of robust evidence for properties. Scientific properties often face elimination, but it is far from clear how often highly robust properties are eliminated. Many properties of neurons for which there was robust evidence in the early 20th century have been retained, such as having a negative transmembrane potential and communicating via synaptic junctions (Guillery 2005). Pessimistic induction reasoning works against RER only if it can be shown that *highly robust* properties have been repeatedly eliminated in the history of science, and it is far from clear whether this is the case.

In sum, none of the objections raised against ER undermine the plausibility of RER. It is a viable and defensible form of scientific realism that deserves to be taken seriously and explored in more detail.

5. Scientific Realism for the Life Sciences

An interesting feature of the debate on scientific realism is that the scientific examples and case studies have almost exclusively been drawn from physics. This is understandable, as it

---

robustness-realist can accept robustness-based IBE that concludes that we are highly justified in believing that X is real, but deny that the IBE from the success of theories to their truth is valid.

is widely assumed that the most mature and explanatorily successful theories and generalizations are found in physics. However, one consequence of this is that accounts of scientific realism often run into problems when applied to the life sciences. For example, Steven French' (2011) discussion of scientific realism in biology merely gestures towards possible ways in which structural realism could be extended to biology in future work. Ladyman and Ross (2007) are more ambitious, and apply their ontic structural realism to the special sciences, but at the cost of reducing all special science entities to patterns that are defined in highly technical information-theoretic terms. This makes their realism completely detached from scientific practice, providing no tools for assessing our degree of justification for the reality of special science entities and properties, and also forcing us to rethink all special science ontologies in terms of structures and patterns.

In contrast, RER directly supports realism in the life sciences, without imposing any kind of ontological revision, and in a way that is continuous with scientific practice. Above I have illustrated this with the example of the neuron, and this is not an isolated or cherry-picked example; the life sciences are full of similar cases. Consider for example mitochondria, cell membranes, pollen or the *Eschericia coli* bacterium. There is extremely robust evidence for each of these entities and many of their properties, and they have been retained in the ontology of biology in spite of radical changes in biological theories.

One related implication of RER is that we may sometimes be more warranted in believing in the reality of entities and properties in the life sciences than in the reality of fundamental physical entities or properties (see also Eronen 2015). For example, when compared to the

neuron, there are relatively few independent ways of measuring, detecting or producing the up quark. Same applies to the recently detected Higgs boson, and for various other elementary particles. In light of the biological examples above, it could turn out that the strongest case studies for scientific realism are not found in physics, as usually has been assumed, but rather in the life sciences. This of course would not mean that these entities and properties of the life sciences are more fundamental than physical entities or properties, but simply that we have more robust evidence for them, and consequently our degree of confidence in their reality is somewhat higher.

As a final remark, it is also possible that ontic structural realism (in the vein of Ladyman and Ross 2007) and RER turn out to be compatible. Ontic structural realism could be seen as a framework for understanding realism in theoretical physics, and for spelling out the metaphysical relationship between special science properties and fundamental physics, while RER could be taken as an account of the science-based ontological commitments in the special sciences. This issue is a topic for future research; in this paper, I hope to have shown that RER is a plausible and defensible form of realism for the life sciences.

References

Cartwright, Nancy. 1983. *How the Laws of Physics Lie.* Oxford: Clarendon Press.

Chakravartty, Anjan. 2007. *A Metaphysics for Scientific Realism*. Cambridge: Cambridge University Press.

Clarke, Steve. 2001. “Defensible Territory for Entity Realism.” *British Journal for the Philosophy of Science* 52: 701-722.

Devitt, Michael. 2005. “Scientific Realism.” In *The Oxford Handbook of Contemporary Philosophy*, ed. Frank Jackson and Michael Smith, 767-791. Oxford: Oxford University Press.

Eronen, Markus. 2015. “Robustness and Reality.” *Synthese* 192: 3961-3977.

French, Steven. 2011. “Shifting to Structures in Physics and Biology: A Prophylactic for Promiscuous Realism.” *Studies in History and Philosophy of Biological and Biomedical Sciences* 42: 164-173.

Guillery, Rainer W. 2005. “Observations of Synaptic Structures: Origins of the Neuron Doctrine and its Current Status.” *Philosophical Transactions of the Royal Society B* 360: 1281-1307.

Hacking, Ian. 1981. “Do We See Through a Microscope?” *Pacific Philosophical Quarterly* 62: 305-322.

Hacking, Ian. 1982. “Experimentation and Scientific Realism.” *Philosophical Topics* 13: 154-172.

Hacking, Ian. 1983. *Representing and intervening. Introductory topics in the philosophy of natural science*. New York: Cambridge University Press.

Hitchcock, Christopher. 1992. “Causal Explanation and Scientific Realism.” *Erkenntnis* 37: 151-178.

Hudson, Robert. 2014. *Seeing Things: The Philosophy of Reliable Observation.* Oxford: Oxfor University Press.

Kuorikoski, Jaakko, Aki Lehtinen, and Caterina Marchionni. 2010. Economic modelling as robustness analysis. *British Journal for the Philosophy of Science* 61: 541–567.

Kuorikoski, Jaakko, and Caterina Marchionni. Forthcoming. “Evidential Diversity and the Triangulation of Phenomena.” *Philosophy of Science.*

Ladyman, James, and Don Ross. 2007. *Every Thing Must Go: Metaphysics Naturalised*. Oxford: Oxford University Press.

Morrisson, Margaret. 1990. “Theory, Intervention and Realism.” *Synthese* 82: 1-22.

Miller, Boaz. Forthcoming. What is Hacking’s Argument for Entity Realism? *Synthese*.

Psillos, Stathis. 1999. *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Raerinne, Jani. 2013. Robustness and sensitivity of biological models. *Philosophical Studies* 166: 285-303.

Scupbach, Jonah. Forthcoming. Robustness Analysis as Explanatory Reasoning. *British Journal for the Philosophy of Science.*

Soler, Lena, Emiliano Trizio, Thomas Nickles, and William C. Wimsatt, eds. 2012. *Characterizing the Robustness of Science: After the Practice Turn in the Philosophy of Science*. Dordrecht: Springer.

Stegenga, Jacob. 2009. Robustness, Discordance, and Relevance. *Philosophy of Science* 76: 650-661.

van Fraassen, Bas. 1980. *The Scientific Image*. Oxford: Oxford University Press.

van Fraassen, Bas. 1985. "Empiricism in the Philosophy of Science." In *Images of Science*, ed. Paul Churchland and Clifford Hooker, 245-308. Chicago: University of Chicago Press.

Wimsatt, William C. 1981. "Robustness, reliability, and overdetermination." In *Scientific Inquiry and the Social Sciences*, ed. M. Brewer and B. Collins, 124-163. San Fransisco: Jossey-Bass.

Wimsatt, William C. (1994). "The ontology of complex systems: levels of organization, perspectives, and causal thickets." *Canadian Journal of Philosophy* S20: 207-274.

Wimsatt, William C. (2007) *Re-Engineering Philosophy for Limited Beings. Piecewise Approximations to Reality*. Cambridge, MA: Harvard University.

Woodward, James (2006). "Some varieties of robustness." *Journal of Economic Methodology* 13: 219–240.

Classical limit and quantum logic

Sebastian Fortin[1,2] – Federico Holik[1,3]

[1] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

[2] Departament of Physics, FCEN, Universidad de Buenos Aires, Argentina.

[3] Instituto de Física de La Plata (IFLP), Argentina.

Abstract

The more common scheme to explain the classical limit of quantum mechanics includes decoherence, which removes from the state the interference terms classically inadmissible since embodying non-Booleanity. In this work we consider the classical limit from a logical viewpoint, as a quantum-to-Boolean transition. The aim is to open the door to a new study based on dynamical logics, that is, logics that change over time. In particular, we appeal to the notion of hybrid logics to describe semiclassical systems. Moreover, we consider systems with many characteristic decoherence times, whose sublattices of properties become distributive at different times.

1. Introduction

In the foundations of physics, the quest of explaining how the laws of classical mechanics arise from the laws of quantum mechanics is known as the *classical limit problem* (Cohen 1989). Generally, this limit is studied for systems that, due to its interaction with the environment, develop a process known as *quantum decoherence* (Schlosshauer 2007). The mathematical description of this phenomenon is usually based on the Schrödinger picture, in which states evolve in time, while observables are taken as constants of motion. Then, projection operators representing physical properties do not evolve in time either. As a result, the structure of the lattice of quantum properties remains the same for all time: the quantum logic associated to the system does not change (Bub 1997).

In this work, we will argue that the description of the lattice of properties in terms of the Schrödinger picture is inadequate for systems undergoing a decoherence process (and thus, it is not useful to describe the *logical* classical limit). We will show that, if the physics of the process represents a transition between quantum to classical mechanics, its logical counterpart should undergo an equivalent transition. Thus, we will propose to study the algebra of the lattice of properties from the perspective of the Heisenberg picture, in which operators representing observables, and their respective projection operators representing physical properties, evolve in time.

From this perspective, we will introduce a novel feature of the classical limit. The study of the time evolution of the projection operators associated to quantum properties in decohering systems opens the way to considering the time evolution of the whole lattice of properties. On this basis, we will study the classical limit from a logical point of view, by describing the manner in which the logical structure of properties associated to observables acquires Boolean features. In other words: we will look for a limit between quantum logic and

Boolean logic and, in this conceptual framework, we will discuss some examples and future perspectives.

2. Observables and Quantum Decoherence

The classical limit problem is usually addressed in terms of the theory of *environment induced decoherence* (EID). This program was developed by the group led by Wojciech Zurek (1982, 1991, 2003), currently at Los Alamos laboratory. According to the Schrödinger picture, a closed quantum system *U*, represented by a state $\hat{\rho}_U(t)$, evolves in time unitarily if no measurements are performed. The system *U* is partitioned into the system of interest *S*, represented by the state $\hat{\rho}_S(t) = \mathrm{Tr}_E\left(\hat{\rho}_U(t)\right)$, and the relevant rest of the world, which is interpreted as the environment *E*, represented by the state $\hat{\rho}_E(t) = \mathrm{Tr}_S\left(\hat{\rho}_U(t)\right)$. The EID approach to decoherence is based on the study of the effects due to the interaction between the quantum system *S*, considered as an open system, and its environment *E*. While *U* evolves in a unitary way, in some typical examples the subsystems may undergo a non-unitary evolution. This allows that, under certain conditions, the state $\hat{\rho}_S(t)$ becomes diagonal after a characteristic decoherence time t_D. In that case, some authors interpret this process as the essence of the classical limit of *S*.

In the framework of the EID approach, quantum decoherence is conceptualized from the point of view of the Schrödinger picture: the phenomenon of decoherence is given in terms of the state evolution. In this representation, the observables associated to the system do not evolve in time. Thus, the commutator between two observables $\hat{O}_1$ and $\hat{O}_2$ stands unchanged during the process. However, decoherence can also be approached to from the viewpoint of the observables of the system.

As it is well known, from the point of view of the properties of the system, the fact that the commutator between two observables vanishes ($\left[\hat{O}_1, \hat{O}_2\right] = 0$) indicates that those

observables are compatible: the corresponding properties can be measured simultaneously. If, on the contrary, the commutator is not zero, $\left[\hat{O}_1, \hat{O}_2\right] \neq 0$, the observables are compatible and the simultaneous measurement of the corresponding properties is not possible. The Schrödinger representation imposes that, if two observables are incompatible at the beginning of the process of decoherence ($t = 0$), then, they will remain incompatible during the entire process, up to its end ($t = t_D$). This fact should call the attention of those who wish to interpret the diagonal state $\hat{\rho}_S(t)$ as a classical state, since in a classical system there are no incompatible observables. Thus, the diagonalization of the reduced state is not sufficient to describe the quantum-to-classical transition of the system.

In the history of decoherence, alternative approaches have been proposed in order to deal with certain problems of EID, in particular, difficulties related to the study of closed systems (Diósi 1987; Milburn 1991; Casati and Chirikov 1995; Polarski and Starobinsky 1996; Adler 2004; Kiefer and Polarski 2009). Among them, we are interested in the self-induced decoherence approach, developed from the physical and philosophical point of view in several papers (Castagnino and Lombardi 2003, 2004, 2005, 2007; Castagnino 2004; Castagnino and Ordóñez 2004; Lombardi and Castagnino 2008; Castagnino and Fortin 2011a). According to the SID approach, a closed quantum system with continuous spectrum may undergo decoherence due to destructive interference, thus reaching a final state that can be interpreted as classical. The central point of this proposal consists in a shift in the perspective: instead of splitting the closed quantum system into "open system" and "environment", the division is traced between *relevant and irrelevant observables*. This mechanism allows us to analyze the time evolution of the mean value of the observables: the vanishing of the interference terms is interpreted as the result of a process of decoherence, which leads to the classical limit.

At this point, it is important to remark that, by means of the commutator between two observables $\hat{O}_1$ and $\hat{O}_2$, it is possible to build a new operator $\hat{C} = i\left[\hat{O}_1, \hat{O}_2\right]$ (Fortin and Vanni 2014). We will interpret this observable as measuring the degree of compatibility between $\hat{O}_1$ and $\hat{O}_2$: if $\hat{C} = 0$, the observables are compatible; if $\hat{C} \neq 0$, they are not. According to quantum mechanics, a closed system evolves unitarily following the Schrödinger equation; since the evolution is unitary, it is impossible that it leads to the following process:

$$\hat{C} \neq 0 \rightarrow \hat{C} = 0$$

In a recent article it has been proved that SID can produce a process of this type in the case of systems with continuous energy spectrum (Fortin and Vanni 2014). Given the incompatible observables $\hat{O}_1$ with core $O_1(\omega, \tilde{\omega})$ and $\hat{O}_2$ with core $O_2(\omega, \tilde{\omega})$, both with continuous spectrum, we can compute the commutator $\hat{C}$ as follows:

$$\hat{C}(t) = i\int_0^\infty\int_0^\infty\int_0^\infty \left(O_1(\omega, \tilde{\omega})O_2(\tilde{\omega}, \omega') - O_2(\omega, \tilde{\omega})O_1(\tilde{\omega}, \omega')\right) d\tilde{\omega}\; e^{i(\omega-\omega')t}\hat{E}_{\omega,\omega'}\, d\omega d\omega'$$

where $\left\{\hat{E}_{\omega,\omega'}\right\}$ is the energy basis of the space of operators. If $O_1(\omega, \tilde{\omega})$ and $O_2(\omega, \tilde{\omega})$ are regular functions, then, by appealing to the Riemann-Lebesgue theorem, it is possible to prove that (see Castagnino and Fortin 2011b)

$$\text{if } \left\langle \hat{C}(t=0) \right\rangle \neq 0 \Rightarrow \lim_{t\to\infty} \left\langle \hat{C}(t) \right\rangle = 0$$

That is, the observable that measures the incompatibility between two observables goes to zero from the observational point of view. This shows that, since the SID approach describes decoherence from the point of view of the mean value of any observable, it turns out to be useful to study the quantum-to-classical transition of $\hat{C}$ (see Fortin and Vanni 2014). As a concrete example, in a Mach-Zender interferometer, if $\hat{O}_1$ is the observable that measures which is the path taken by the photon, and $\hat{O}_2$ is the observable associated to the visibility of interference, then, $\hat{C}$ can be conceived as the tool to measure how compatible those

observables are. In the lab, there are different observables associated with the degree of classicality; for example, the contrast of the interference fringes in the double slit experiment. When the experiment is performed and decoherence occurs, it is reasonable to expect that at the beginning $\hat{C} \neq 0$, but then, after the decoherence time, the system reaches the classical limit with $\hat{C} = 0$. And it is also expected that, in that limit, the interference fringes will accordingly vanish. Moreover, in an experiment with slow and controlled decoherence, it could be possible to measure the evolution of the observable $\hat{C}$.

EID and SID are not the only ways to account for non-unitary evolutions. A strategy to transform the unitary evolution of a closed system into a non-unitary evolution has been proposed in the cosmological context. Kiefer and Polarski (2009) adopted the Heisenberg picture for the study of the decoherence process of the universe. According to this perspective, the state $\hat{\rho}$ stands constant while the observables $\hat{O}(t)$ change in time. In this way, the observable associated to the commutator of two observables becomes a function of time, $\hat{C}(t)$. This approach allows us to study the commutator of two observables for cosmological problems. In particular, according to the *inflation model*, there was an accelerated phase of the early universe called inflation; the whole structure of the universe can be traced back to the primordial fluctuations in the *inflaton* field (Kolb and Turner 1990; Mukahnov 2005; Peacock 1990). Because of the expansion of the universe, inflaton fluctuations must be described by a time-dependent Hamiltonian:

$$\hat{H}(\eta) = \frac{1}{2}\int dk^3 \left[k\left(\hat{a}(k)\hat{a}^{\dagger}(k) + \hat{a}^{\dagger}(-k)\hat{a}(-k)\right) + i\frac{a'}{a}\left(\hat{a}^{\dagger}(k)\hat{a}^{\dagger}(k) + \hat{a}(-k)\hat{a}(-k)\right)\right]$$

where η is the conformal time, $\hat{a}(k)$, $\hat{a}^{\dagger}(k)$ are the annihilation operator and the creator operator respectively, and a is the scale factor of the universe. These three last elements are time dependent, and this is the reason why the Hamiltonian $\hat{H}(\eta)$ is not constant in time.

Under these conditions, it is possible to compute the commutator between the operators of position $\hat{y}(\eta)$ and momentum $\hat{p}(\eta)$ (see Kiefer and Polarski 2009):

$$\left[\hat{y}(0), \hat{p}(0)\right] \neq 0 \Rightarrow \lim_{\eta \to \infty} \left[\hat{y}(\eta), \hat{p}(\eta)\right] = 0$$

In other words, the evolution of the commutator between the operators of position and momentum shows that, under certain conditions, it vanishes for times longer than the decoherence time.

Finally, it is important to mention that the approach to decoherence based in non-Hermitian Hamiltonians was also applied to the study of the time evolution of the commutators (Fortin, Holik and Vanni 2016).

3. The Logical Perspective

As it is well known, any physical observable of a quantum system can be represented in a mathematical way as a self-adjoint operator on a Hilbert space (Ballentine 1990). The *spectral theorem* states that any self-adjoint operator $\hat{A}$ can be represented by its *projective measure* $M_A(...)$ (Reed and Simon 1972; Rèdei 1998; Lacki 2000). A projective measure assigns a projection operator to each Borel set of the real line: given the interval $I(a,b)$, $M_A(I)$ is a projection operator. This mathematical fact was interpreted by Birkhoff and von Neumann (1936) as follows. The projector $M_A(I)$ represents the empirical proposition: "the value of the observable represented by $\hat{A}$ lies in the interval I". The truth value of this proposition can be obtained experimentally by means of a yes-no test: that truth value can be tested in any particular run of the experiment, and the quantum state assigns a probability to it.

These formal aspects of quantum theory constitute the elemental bricks out of which the entire building of its rigorous formulation is erected; this task was achieved by von Neumann (1932) in his famous *Mathematical Foundations of Quantum Mechanics*. Importantly enough,

the same kind of analysis can be performed for classical probabilistic theories, and further research showed that this approach can be extended to quantum field theory and quantum statistical mechanics. The algebraic structure of the quantum mechanical propositions was called *quantum logic* after the famous paper by Birkhoff and von Neumann (1936). As it is well known, those propositions can be endowed with an *orthomodular lattice* structure (Kalmbach 1983). Additionally, a solid axiomatic foundation for quantum mechanics can be used to explain in an operational way many important features of the Hilbert space formalism (Varadarajan 1968; Stubbe and Van Steirteghem 2007; see also Holik et al. 2013, 2014, 2015 for more recent developments, and for the relationship between the quantum-logical approach and quantum probability theory). But the feature relevant to our discussion is that the logic associated to all varieties of quantum theories is not Boolean, due to the fact that it is not distributive. This implies a very deep structural difference between classical and quantum theories.

Quantum states are, in its formal essence, measures that assign probabilities to all the different empirical propositions. For example, if we want to know the probability of observing the value of the observable $\hat{A}$ in the interval I, given that the system is prepared in the state $\hat{\rho}$, the Born rule states that this quantity is given by $Tr[\rho M_A(I)]$. According to the traditional Schrödinger picture, unitary evolutions induce time transformations between states. But, according to the Heisenberg picture, observables are transformed, and this transformation induces an action on their respective spectral measures. This in turn implies that the actual properties (i.e., those involved in propositions whose truth is endowed with probability equal to one) also evolve in time. In other words, unitary time evolutions are represented by *automorphisms* on the quantum logic (they are just "rotations" in the projective geometry of the Hilbert space). More general evolutions (such as the non-unitary evolutions associated to

measurements or to decoherence processes) are represented by Kraus operators, and also induce concomitant maps on the quantum logic.

But although all possible kinds of time evolutions can be described in the rigorous approach to quantum theory, decoherence poses a conceptual problem in the following sense. Let us suppose that we start with a system that is completely quantum, with its associated orthomodular lattice of projection operators. If the system undergoes a classical limit process, the lattice associated to the final stage should be classical (i.e., Boolean). Therefore, if we want to describe faithfully the classical limit, we should have at hand a time ordered family of logics, starting from a quantum one, and ending up with a classical one. This is the problem that we are going to address in the next section. Transitions between logics were studied (see, for example, Aerts et al. 1993), but not in relation to decoherence and the classical limit. In the present work, we are interested in the philosophical implications of assuming a non-unitary time evolution to induce a continuous family of logics to describe the process of the classical limit. As we will see, this perspective leads to a better understanding of this physical process, and is also useful to cope with hybrid systems.

4. The Classical Limit from the Logical Point of View

In order to be able to describe the classical limit from a logical point of view, let us consider a quantum system that evolves in a non-unitary way, and a set of relevant observables represented by self-adjoint operators, $O = \left\{ \hat{O}_1, \hat{O}_2, \hat{O}_3, ..., \hat{O}_N \right\}$. Let us also consider the algebra $V(0)$ generated by O at time $t = 0$. We also assume that some of the observables of O are incompatible: for some *i* and *j*, we initially have $\left[\hat{O}_i, \hat{O}_j \right] \neq 0$. In a system with these features, the condition for the classical limit –according to the Heisenberg picture– is given by the following evolution:

$$\forall i, j, \quad \left[\hat{O}_i(0), \hat{O}_j(0) \right] \neq 0 \rightarrow \left[\hat{O}_i(t_D), \hat{O}_j(t_D) \right] = 0$$

As time passes, the evolving operators generate a family of algebras $V(t)$. The final algebra, $V(t_D)$ is a Boolean algebra since, if the classical limit is reached successfully, the final set of generating operators will be a set of pairwise commutative operators. That is: initially incompatible observables become compatible after the decoherence time. The algebras $V(t)$ have associated orthomodular lattices $L_{V(t)}$: the classical limit is expressed by the fact that, while $L_{V(0)}$ is a non-distributive lattice of projectors, $L_{V(t_D)}$ is a Boolean one. In this way, we obtain an adequate description of the logical evolution of a quantum system.

4.1 Semiclassical systems from the logical point of view.

The condition that imposes that all observables of the system must be commutative is equivalent to that of the diagonalization of the state operator, and it is necessary in the case of quantum systems that become completely classical. Notwithstanding, if this condition is strictly applied to any case of classical limit, it leaves no room for the description of the majority of everyday systems, some of which of great importance, such as transistors or squids (Clarke and Braginski 2004). As an example, let us suppose that we go to an electronics store to buy a transistor. The salesman will first find its location in the shelves, and then will take it with his hand in order to put it in a bag and, finally, to give it to us. From this point of view and for all practical purposes, the transistor *behaves classically*: it is an object that can be located in space an time, and can be manipulated by classical means. However, when connected to a circuit, well-known quantum effects of our interest take place on it; for example, consider the tunnel effect of the electrons inside it. This means that a transistor is an object such that some of its observables behave classically, while some others behave in a quantum way: physicists refer to objects of this kind as *semiclassical*.

Our approach of the classical limit allows us to account for these cases. In the semiclassical situation, instead of the above strong condition, the condition turns out to be:

$$\exists i, j, \quad \left[\hat{O}_i(0), \hat{O}_j(0)\right] \neq 0 \rightarrow \left[\hat{O}_i(t_D), \hat{O}_j(t_D)\right] = 0$$

In other words, there are some observables that begin as incompatible and become compatible through the evolution. But there may be also observables that are incompatible at the beginning, and remain incompatible after the decoherence time. From a logical viewpoint, this implies that the lattices of properties associated to this kind of systems are *hybrid* lattices.

The focus on hybrid lattices is of particular importance, because it is reasonable to suppose that, if successfully developed, quantum computers will be semiclassical systems in their very nature, represented by hybrid lattices. This is manifested by the fact that some relevant quantum algorithms possess classical and quantum elements in the process of computation (see, for example, Shor 1997). Thus, a hybrid logic might be useful not only to describe the logical architecture of a quantum computer in a conceptual way, but also to cope with the problems related to decoherence.

4.2 Transitions using many steps

Up to this point we have considered quantum systems that become classical after a decoherence time t_D; in this way, we explained the transition from a quantum logic to a Boolean logic. But we have not explored in detail the intermediate steps of this transition. One way to do this is to consider systems with several characteristic times.

There are a number of examples of physical systems that reach the classical limit in several stages. From the point of view of the state operator, this means that its different non-diagonal components vanish at different characteristic times (Fortin, Holik and Vanni 2016). A concrete example of such a system is that of a harmonic oscillator embedded in a bath of oscillators (Castagnino and Fortin 2012). In this case, the compatibility condition between different observables is fulfilled at different times as follows:

$$\left[\hat{O}_1(0),\hat{O}_2(0)\right]\neq 0\rightarrow\left[\hat{O}_1(t_\alpha),\hat{O}_2(t_\alpha)\right]=0$$
$$\left[\hat{O}_1(0),\hat{O}_3(0)\right]\neq 0\rightarrow\left[\hat{O}_1(t_\beta),\hat{O}_3(t_\beta)\right]=0$$
$$\cdots$$
$$\forall i,j,\quad \left[\hat{O}_i(0),\hat{O}_j(0)\right]\neq 0\rightarrow\left[\hat{O}_i(t_D),\hat{O}_j(t_D)\right]=0$$

To put it into words: among all the observables that are incompatible at the beginning of the process, some become compatible at time t_α, others become compatible at time t_β, and so on. If the classical limit is reached, at the end of the process all the observables will commute with each other. In the logical language of lattices introduced above, this many-step process can be described by stating that the different parts of the evolving lattice will become distributive at different times.

5. Conclusions

Since the very beginnings of quantum mechanics, many attempts have been made to recover the laws of classical physics from quantum mechanics through a classical limiting process. This classical limit must do the job of turning a quantum system –described by a quantum logic at $t=0-$ into a classical system –described by a Boolean logic at the end of the limiting process, at t_D in the case of decoherence. The dynamical characteristics of the quantum-to-classical transition were extensively studied in the physical literature. However, from a logical perspective, the quantum-to-Boolean transition was usually merely understood as a jump from a quantum logic at $t=0$ to a Boolean one at t_D. Accordingly, researchers did not pay attention to the logical structures associated to the system in times belonging to the interval $(0,t_D)$. As an example of this non-trivial logical structure, we presented physical systems with different characteristic times, which, as a consequence, reach the classical limit in many steps. This shows that the study of the logical features of intermediate times in a quantum-to-classical limiting process may exhibit a rich and non-trivial dynamical structure.

In this work, we described the decoherence process by appealing to the Heisenberg's picture. We argued that it is the proper framework for studying the quantum-to-Boolean transition. With this useful tool, we analyzed the transition in three different cases: (i) logical classical limit in systems with one characteristic time; (ii) systems that change from a quantum logic to a hybrid semiclassical logic; and (iii) systems with many characteristic decoherence times, whose sublattices become distributive at different times. The description of the classical limit presented in this short work does not claim to be exhaustive or complete. But it intends to be the kickoff for the study of a largely unexplored area of the logical structure of quantum systems. Studies of this kind might be of great help in the understanding of the new technologies associated to quantum computers (which involve hybrid logics) and to general quantum information processing tasks.

6. References

Adler, Stephen 2004. *Quantum Theory as an Emergent Phenomenon*. Cambridge: Cambridge University Press.

Aerts, Diederik, Thomas Durt, and Bruno Van Bogaert 1993. "Quantum Probability, the Classical Limit and Nonlocality." In *Symposium on the Foundations of Modern Physics 1992: The Copenhagen Interpretation and Wolfgang Pauli*, ed. K. V. Laurikainen and C. Montonen, 35-56. Singapore: World Scientific.

Ballentine, Leslie 1990. *Quantum Mechanics*. New York: Prentice Hall.

Birkhoff, George, and John von Neumann 1936. "The Logic of Quantum Mechanics." *Annals of mathematics* 37: 823-843.

Bub, Jeffrey 1997. *Interpreting the Quantum World*. Cambridge: Cambridge University Press.

Casati, Giulio, and Boris Chirikov 1995. "Quantum Chaos: Unexpected Complexity." *Physica D* 86: 220-237.

Castagnino, Mario 2004. "The Classical-Statistical Limit of Quantum Mechanics." *Physica A* 335: 511-517.

Castagnino, Mario, and Sebastian Fortin 2011a. "New Bases for a General Definition of the Moving Preferred Basis." *Modern Physics Letters A* 26: 2365-2373.

—— 2011b. "Formal Features of a General Theoretical Framework for Decoherence in Open and Closed Systems." *International Journal of Theoretical Physics* 52, 1379-1398.

—— 2012. "Non-Hermitian Hamiltonians in Decoherence and Equilibrium Theory." *Journal of Physics A* 45: 444009.

Castagnino, Mario, and Olimpia Lombardi 2003. "The Self-Induced Approach to Decoherence in Cosmology." *International Journal of Theoretical Physics* 42: 1281-1299.

—— 2004. "Self-Induced Decoherence: A New Approach." *Studies in History and Philosophy of Modern Physics* 35: 73-107.

—— 2005. "Self-Induced Decoherence and the Classical Limit of Quantum Mechanics." *Philosophy of Science* 72: 764-776.

—— 2007). "Non-Integrability and Mixing in Quantum Systems: On the Way to Quantum Chaos." *Studies in History and Philosophy of Modern Physics* 38: 482-513.

Castagnino, Mario, Adolfo Ordóñez 2004. "Algebraic Formulation of Quantum Decoherence." *International Journal of Theoretical Physics* 43: 695-719.

Clarke, John, and Alex Braginski 2004. *The SQUID Handbook: Fundamentals and Technology of SQUIDs and SQUID Systems, Volume I*. Weinheim: Wiley-VCH.

Cohen, David 1989. *An Introduction to Hilbert Space and Quantum Logic*. Berlin: Springer-Verlag.

Diósi, Lajos 1987. "A Universal Master Equation for the Gravitational Violation of Quantum Mechanics." *Physics Letters A* 120: 377-381.

Fortin, Sebastian, and Leonardo Vanni 2014. "Quantum Decoherence: a Logical Perspective." *Foundations of Physics* 44: 1258-1268.

Fortin, Sebastian, Federico Holik, and Leonardo Vanni 2016. "Non-Unitary Evolution of Quantum Logics." *Springer Proceedings in Physics*, forthcoming.

Holik, Federico, and Angelo Plastino 2015. "Quantum Mechanics: A New Turn in Probability Theory." In *Contemporary Research in Quantum Systems*, ed. Z. Ezziane, 399-414, New York: Nova Publishers.

Holik, Federico, César Massri, Angelo Plastino, and Leandro Zuberman 2013. "On the Lattice Structure of Probability Spaces in Quantum Mechanics." *International Journal of Theoretical Physics* 52: 1836-1876.

Holik, Federico, Angelo Plastino, and Manuel Sáenz 2014. "A Discussion on the Origin of Quantum Probabilities." *Annals of Physics* 340: 293-310.

Kalmbach, Gudrun 1983. *Orthomodular Lattices*. San Diego: Academic Press.

Kiefer, Claus, and David Polarski 2009. "Why Do Cosmological Perturbations Look Classical to Us?" *Advanced Science Letters* 2: 164-173.

Kolb, Edward, and Michael Turner 1990. *The Early Universe*. Reading MA: Addison-Wesley.

Lacki, Jan 2000. "The Early Axiomatizations of Quantum Mechanics: Jordan, von Neumann and the Continuation of Hilbert's Program." *Archive for History of Exact Sciences* 54: 279-318.

Lombardi, Olimpia, and Mario Castagnino 2008. "A Modal-Hamiltonian Interpretation of Quantum Mechanics." *Studies in History and Philosophy of Science* 39: 380-443.

Milburn, Gerard 1991. "Intrinsic Decoherence in Quantum Mechanics." *Physical Review A* 44: 5401-5406.

Mukhanov, Viatcheslav 2005. *Physical Foundations of Cosmology*. Cambridge: Cambridge University Press.

Peacock, John 1999. *Cosmological Physics*. Cambridge: Cambridge University Press.

Polarski, David, and Aleksei Starobinsky 1996. "Semiclassicality and Decoherence of Cosmological Perturbations." *Classical and Quantum Gravity* 13: 377-392.

Rèdei, Mikos 1998. *Quantum Logic in Algebraic Approach*. Dordrecht: Kluwer Academic Publishers.

Reed, Michel, and Barry Simon 1972. *Methods of Modern Mathematical Physics I: Functional Analysis*. New York: Academic Press.

Schlosshauer, Maximilian 2007. *Decoherence and the Quantum-to-Classical Transition*. Berlin: Springer.

Shor, Peter 1997. "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithm on a Quantum Computer." *SIAM Journal on Computing* 26: 1484-1509.

Stubbe, Isar, and Bart Van Steirteghem 2007. "Propositional Systems, Hilbert Lattices and Generalized Hilbert Spaces." In *Handbook of Quantum Logic Quantum Structures:*

Quantum Structures, ed. K. Engesser, D.M. Gabbay and D. Lehmann, 477-523. Amsterdam: Elsevier.

Varadarajan, Veeravalli 1968. *Geometry of Quantum Theory I*. Princeton: van Nostrand.

von Neumann, John 1932. *Mathematische Grundlagen der Quantenmechanik*. Heidelberg: University Press.

Zurek, Wojciech 1982. "Environment-Induced Superselection Rules." *Physical Review D* 26: 1862-1880.

—— 1991. "Decoherence and the Transition from Quantum to Classical." *Physics Today* 44: 36-44.

—— 2003. "Decoherence, Einselection, and the Quantum Origins of the Classical." *Reviews of Modern Physics* 75: 715-775.

Word count: 4298

Abstract

Experiments demonstrating entanglement swapping have been alleged to challenge realism about entanglement. Seevinck (2006) claims that entanglement "cannot be considered ontologically robust" while Healey (2012) claims that entanglement swapping "undermines the idea that ascribing an entangled state to quantum systems is a way of representing some new, non-classical, physical relation between them." My aim in this paper is to show that realism is not *threatened* by the possibility of entanglement swapping, but rather, it should be *informed* by the phenomenon. I argue—expanding the argument of Timpson and Brown (2010)—that ordinary entanglement swapping cases present no new challenges for the realist. With respect to the delayed-choice variant discussed by Healey, I claim that there are two options available to the realist: (a) deny these are cases of genuine swapping (following Egg (2013)) or (b) allow for existence of entanglement between timelike separated regions. This latter option, while radical, is not incoherent and has been suggested in quite different contexts. While I stop short of claiming that the realist *must* take this option, doing so allows one to avoid certain costs associated with Egg's "orthodox" account. I conclude by noting several important implication of entanglement swapping for how one thinks of entanglement generally.

Swapping Something Real

April 6, 2016

1 Introduction

The phenomenon of quantum entanglement has been taken to have broad metaphysical implications.[1] Such implications presuppose a broadly realist view of entanglement, one that recognizes a genuine physical relation between the subsystems that compose an entangled system. This entanglement relation, in turn, is used to explain the sorts of non-local correlations found in the measurement results of EPR-B[2] and related experiments. These correlations are "non-local" in that they hold between distant measurement events that occur at the same time—i.e., at spacelike separation.

Recent experiments involving "entanglement swapping," threaten to complicate our typical understanding of entanglement. Some have even suggested that these experiments threaten to undermine the realist position altogether. Below I will argue that this isn't the case. However, entanglement swapping is not without important implications for the realist. Indeed, I claim that delayed-choice entanglement swapping gives us reason to consider extending

[1] Ladyman and Ross claim that "entanglement as described by QM teaches us that Humean supervenience is false, and that all the properties of fundamental physics seem to be extrinsic to individual objects" (2007, 151). A similar claim is made by Esfeld (2004), who claims that entanglement recommends a "metaphysics of relations." Quantum entanglement also pays a critical role in Schaffer's (2010) defense of monism, the view that there is ultimately only one object: the entire universe.

[2] I use "EPR-B" to refer to variations of the experimental arrangement due to Einstein et al. (1935) and extended by Bohm (1951). The variations most relevant in what follows will be those involving photon pairs with entangled polarizations.

entanglement into the temporal dimension. By allowing for timelike entanglement, the realist is able provide a unified account of a variety of experimental results. Even if one rejects this radical suggestion, ordinary cases of entanglement swapping alone require revising widely-held views about the nature of entanglement.

2 Preliminaries

Quantum theory doesn't wear its metaphysics on its sleeve. Different interpretations of quantum theory radically diverge on what (if anything) it tells us about the world. Accordingly, it is impossible to undertake our investigation without making some interpretative assumptions. That said, many of the issues here cross-cut interpretations and I hope to remain as neutral as possible between the various realist interpretations. I begin with the orthodox view of how entanglement arises in the formalism of (ordinary, non-relativistic) quantum mechanics.

2.1 Nonseperable quantum states

Quantum mechanics allows for nonseparable quantum states. To keep matters as simple as possible, consider two particles, 1 and 2, each of which can be assigned a pure quantum state. The standard approach represents the quantum state of each particle with a vector (ray) $|\psi\rangle$ in a Hilbert space $\mathcal{H}$. The quantum states of two systems 1,2, then, correspond to vectors $|\psi\rangle, |\phi\rangle$ in Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$, respectively. The joint state of the system they compose is represented by the vector $|\Psi\rangle$ in the tensor product Hilbert space $\mathcal{H}_{12} = \mathcal{H}_1 \otimes \mathcal{H}_2$. If the state vector $|\Psi\rangle$ in $\mathcal{H}_{12}$ can be expressed as a product of vectors $|\psi\rangle, |\phi\rangle$ in Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$, then $|\Psi\rangle$ is *separable*. But, in general, a vector in $\mathcal{H}_{12}$ cannot be expressed in the form $|\Psi\rangle = |\psi\rangle \otimes |\phi\rangle$, with $|\psi\rangle \in \mathcal{H}_1$ and $|\phi\rangle \in \mathcal{H}_2$. Such states are called *nonseparable quantum states*.

On the standard view, entanglement occurs when distinct physical systems are attributed nonseparable quantum states. Thus, if two photons 1,2 are prepared in the nonseperable joint polarization state $|\Psi^-\rangle = \frac{1}{\sqrt{2}}(|HV\rangle - |VH\rangle)$,

they (or their quantum states) are mutually entangled. While this standard view of entanglement has been criticized (Ghirardi et al. 2002; Ladyman et al. 2013), all of the cases considered below will count as entangled on any suitable definition. Accordingly, I will bracket worries about the precise formulation of entanglement in the quantum formalism and simply assume the standard account for ease of exposition.

2.2 Entanglement realism

In order to say more about entangled systems, we must go beyond the formalism of quantum theory. What is the significance of ascribing entangled states to a set of physical systems?

In what follows, I will be concerned with views that accord the quantum state a *descriptive* role. Thus, when we attribute entangled states to composite systems, that tells us something about the relation between the physical subsystems in question. I will aim to remain as neutral as possible about the nature of this relation. The following are two possible views about the nature of this relation:

Action at a Distance: On this view, distant entangled subsystems are capable of having an immediate and unmediated causal influence on each other.

Ontological Holism: On this view, a compound entangled system is viewed as a nonseparable whole, which is irreducible to the subsystems it comprises.

Other variations of these views are possible as well. Some maintain that entangled systems are connected by a new *non-supervenient relation* while others speak of non-local influence that fails to be genuinely causal. It is not my aim here to adopt any particular approach to the metaphysics of entanglement. Rather, what will be at issue is the following thesis:

Entanglement Realism: Entangled systems bear a genuine physical relation to one another—one that is constitutive of their mutual entanglement.

Entanglement realism cross-cuts interpretations of quantum theory. Broadly "anti-realist" interpretations such as instrumentalism and other epistemic views will deny the thesis, but so will some characteristically "realist" views as well. First, consider an instrumentalist that views the quantum state epistemically. On this view, the assignment of a non-separable quantum state is a way of summarizing our information about the system. While ascribing such a quantum state allows us to predict non-local correlations, this view stops short of recognizing a physical entanglement relation between the particles themselves (if there are such things). Second, consider a Bohmian who takes the motion of particles to be fundamental and understands the wavefunction as a law-like feature of how particles move. On such a view, an entangled quantum state does not support the existence of a new physical relation between particles, but only describes/guides the motion of the particles so as to generate non-local correlations. There is not space to discuss all possible interpretations of quantum theory and their relation to entanglement realism, nor is this the appropriate place to debate the merits of the view. Instead, I'll conclude this section with two remarks intended to clarify the position.

First, whether an interpretation endorses entanglement realism depends solely on whether there is a physical relation R that can be attributed to a compound physical system in virtue of it being ascribed a nonseparable quantum state; being a "realist" interpretation isn't sufficient (though it may be necessary). Second, as with other forms of realism, the primary motivation for entanglement realism is explanatory. However the entanglement relation is understood, it should enable robust explanations of non-local correlations in measurement results. Relatedly, views the deny entanglement realism do so at the potential cost of being unable to adequately explain non-local correlations. Thus, there is at least some reason (*ceteris paribus*) to prefer interpretations of quantum theory that countenance entanglement realism.

3 Entanglement swapping

The experiments that motivate entanglement across time make use of the technique of *entanglement swapping*. Entanglement swapping is a relatively

recent phenomena, and as a result has received relatively little consideration by philosophers. A simple experimental arrangement is depicted below (figure 1). Consider two sources that each produce a pair of photons in the state $|\psi^-\rangle = \frac{1}{\sqrt{2}}(|HV\rangle - |VH\rangle)$. One source produces the entangled pair (1,2) and the other produces (3,4). Initially, the quantum state of the four-particle system is simply the product of two pair states $|\Psi\rangle = |\psi^-\rangle_{12} \otimes |\psi^-\rangle_{34}$. This state is separable into the states $|\psi^-\rangle_{12}$ and $|\psi^-\rangle_{34}$, each of which is an entangled two-photon state. Accordingly, entanglement realist would initially recognize two distinct entanglement relations—R_{12} and R_{34}—but no such relations between the pairs or between photons from different pairs.

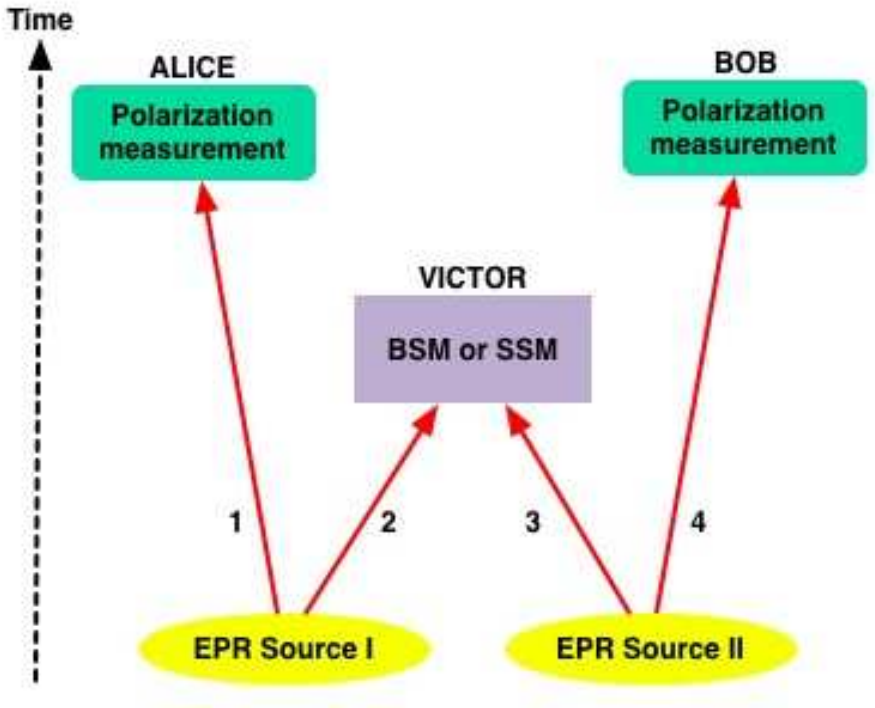


Figure 1: Entanglement Swapping Configuration

The outermost particles are sent off to polarization detectors at Alice and Bob. The inner particles are sent to a common location, Victor, which contains a switchable Bell-state analyzer. When switched on, a Bell-state measurement (BSM) is performed, which has the effect of projecting the indent particles into

one of the four entangled Bell-states.[3] Otherwise, a separate state measurement (SSM) is performed. If the analyzer is off and the particles are measured separately, then, as expected, correlations are found between (1,2)and (3,4) as in an ordinary EPR-B experiment.

If the analyzer is on, however, particles 2 and 3 are projected into one of the entangled Bell-states and, as a result, the remaining particles 1 and 4 are projected into an entangled Bell-state as well. This can be seen by writing the initial four-particle state in the basis given by the Bell-states of (1,4):

$$|\Psi\rangle = \frac{1}{2}[|\psi^+\rangle_{14}|\psi^+\rangle_{23} - |\psi^-\rangle_{14}|\psi^-\rangle_{23} - |\phi^+\rangle_{14}|\phi^+\rangle_{23} - |\phi^-\rangle_{14}|\phi^-\rangle_{23}]. \quad (1)$$

Given this expression of $|\Psi\rangle$, we can see that if a BSM is performed at Victor with the result $|\psi^+\rangle_{23}$, then the remaining particles are projected into the state $|\psi^+\rangle_{14}$, and similarly for the other Bell states. Crucially, regardless of the outcome of the BSM at Victor, photons 1 and 4 become entangled as a result. This is the case despite the fact that they have never interacted.

At least some have taken this case to problematize entanglement realism:

> that one cannot think of entanglement as a property [which] has some ontological robustness can already be seen using the following weaker requirement: anything which is ontologically robust can, without interaction, not be mixed away, nor swapped to another object, nor flowed irretrievably away into some environment. Precisely these features are possible in the case of entanglement and thus even the weaker requirement for ontological robustness does not hold. (Seevinck 2006, 1582)

The intuition underlying this challenge is that something real would require a genuine "interaction" to be altered, but entanglement swapping allows us to move the entanglement around without such an interaction. But is it

[3]For polarization measured along the H/V axis these are:

$$|\psi^\pm\rangle = \frac{1}{\sqrt{2}}[|H\rangle|V\rangle \pm |V\rangle|H\rangle], \quad |\phi^\pm\rangle = \frac{1}{\sqrt{2}}[|H\rangle|H\rangle \pm |V\rangle|V\rangle].$$

really the case that there is no interaction responsible for the swapping? After all, particles 2 and 3 are directly affected by the BSM performed at Victor. However, nothing is done directly to the remaining particles 1 and 4, and it is these that become entangled, so perhaps there is something amiss. Indeed, it is puzzling how exactly 1 and 4 become entangled remotely and instantaneously, but this is simply the original problem of entanglement in another form.

According to the realist who posits non-local influence, the ordinary EPR case is already one in which the measurement of a spacelike separated system affects the properties of a system entangled with it. If, however, we have some way of understanding such influence in terms of a physical entanglement relation, then presumably that relation can do the necessary work needed to account for entanglement swapping. In the case of a SSM at Victor, measurements of 1 and 4 will display correlations with the results obtained at Victor. In the case of a BSM, there are not simple correlations between the measurement at Victor and those at Alice and Bob, but rather, a more complex pattern of relations best accounted for by attributing an entangled Bell-state to the joint (1,4) system.

Timpson and Brown (2010) agree that entanglement swapping fails to provide a convincing case against entanglement realism. They suggest an analogy with gravity in Newtonian physics to illustrate:

> We do not think that the relative distance between two planets in Newtonian physics is not a genuine feature of reality because of the action-at-a-distance of the gravitational interaction. (Timpson and Brown 2010, 317)

I take the suggestion to be the following. Just as the Newtonian might seek to explain a pattern in the motion of two planets by appeal to a pattern in the motion of two other planets connected to them by an instantaneous gravitational influence, entanglement relations could provide a similar connection between the pairs of particles between which entanglement is swapped. Note, however, that adopting such a view requires a somewhat broader understanding of action at a distance than is ordinary supposed. Standard formulations focus on the *intrinsic properties* of systems. For instance, in his Stanford

Encyclopedia article on the topic, Berkovitz defines action at a distance as:

> a phenomenon in which a change in intrinsic properties of one system induces a change in the intrinsic properties of a distant system without there being a process that carries this influence contiguously in space and time. (Berkovitz 2016).[4]

To account for entanglement swapping in the manner above, the proponent of action at a distance must allow that the relational properties of particles (i.e., their entanglement relations) can influence the relational properties of the particles with which they are entangled.

How significant of a revision is this? One could claim, along holist lines, that entanglement is an intrinsic property of the *compound* system, in which case the ordinary version of action at a distance perhaps could be preserved. At least on the "orthodox" understanding of quantum mechanics, however, there is no clear basis for attributing an intrinsic property to a bipartite system on the basis of entanglement between its constituents. The extension from intrinsic properties to relations is certainly in keeping with the spirit of action at a distance, as the analogy with Newtonian gravity suggests, but it is a significant change none the less. Entanglement must now be understood as capable of spreading new entanglement relations, which is no doubt an interesting result.

Similar revisions are required for the holist to account for entanglement swapping. When the photons are created there are two pairs of mutually entangled particles. Hence, the holist would recognize (fundamentally) two two-photon wholes, (1,2) and (3,4), that are spreading out spatially with time. Victor's measurement is performed on both wholes and immediately alters both. If a SSM is performed, each two-photon system dissolves leaving photons 1 and 4 to be detected later. If a BSM is performed, again each two-photon system is changed, but in a way that the new wholes (2,3) and (1,4) are formed. Thus, the holist must allow that certain measurements are capable of generating new wholes out the parts of the original ones. Again this is does seem to mark an important revision in the view, but not one that creates any

[4]This is the broader of two definitions given by Berkovitz, both of which contain a reference to intrinsic properties.

obvious problems.

Before moving to the next section, it is worth noting that entanglement swapping is not a mere philosophical curiosity, but is part of an active research program in physics with numerous practical applications, including: constructing a quantum telephone exchange, speeding up the distribution of entanglement, correcting errors in Bell states, preparing entangled states of a higher number of particles, and secret sharing of classical information (Bouwmeester et al. 2000). This makes its dismissal or reinterpretation difficult to motivate from a realist perspective. A key tenant in realist thinking recommends endorsing those parts of scientific theory that facilitate predictive and technological successes such as these.

4 Delayed-choice entanglement swapping

The revision to our understanding of entanglement required by entanglement swapping cases like that depicted in figure 1 is consistent with the central ideas of action at a distance or ontological holism. Entanglement swapping with delayed-choice, by contrast, threatens to undermine such notions completely.

> The delayed-choice entanglement-swapping experiment reinforces the lesson that quantum states are neither descriptions nor representations of physical reality. In particular, it undermines the idea that ascribing an entangled state to quantum systems is a way of representing some new, non-classical, physical relation between them. (Healey 2012, 31)

The idea of delayed-choice entanglement swapping was first proposed by Peres (2000). We begin with two entangled systems as in the ordinary case, but rather than have Victor preform his measurement prior to Alice and Bob, we delay particles 2 and 3 so that Victor can perform his measurement *after* his colleagues. Because the explanation of the collapse of equation 1 into entangled Bell-states of (2,3) and (1,4) didn't specify any times, quantum mechanics suggests that the same results would obtain. In particular, when

Victor successfully performs a BSM, entanglement will be swapped to 1 and 4.

And, in fact, these results seem to have been confirmed by an experiment conducted by Ma et al. (2012) depicted below (Figure 2). We begin as before: two pairs of entangled photons (1,2) and (3,4) are produced by two EPR sources in the state $|\psi^-\rangle_{12} \otimes |\psi^-\rangle_{34}$. At this point the photons 1 and 2 are mutually entangled, as are 3 and 4, but the 4-particle state is separable, and hence there is no entanglement across the two pairs. Alice and Bob each perform a polarization measurement of their photon (1 and 4, respectively) along one of three freely-chosen axes ($|H\rangle/|V\rangle, |R\rangle/|L\rangle, |+\rangle/|-\rangle$) and the data from these measurements are saved for later analysis. Particles 2 and 3, meanwhile, enter an optical delay, and only reach Victor at time M_V, nearly 500ns after M_A and M_B, the times at which Alice and Bob perform their measurements.

As before, Victor "chooses" between performing a Bell-state measurement (BSM) or separate state measurement (SSM) on (2,3). In the actual experiment, the switchable Bell-state analyzer was linked to a quantum random number generator which determined the measurement (BSM or SSM) to be performed. The photons 2 and 3 are projected into either an entangled state ($|\phi^+\rangle_{23}$ or $|\phi^-\rangle_{23}$) if BSM is performed or a separable state in the case of SSM. When Victor's results are compared with those of Alice and Bob, they are found to be consistent with ascribing an entangled state to photons 1 and 4 ($|\phi^+\rangle_{14}$ or $|\phi^-\rangle_{14}$) when BSM is performed and a separable state otherwise. Thus, it seems that entanglement has been swapped to particles (1,4) *after* they have already been detected (at M_V)!

This is puzzling to the entanglement realist. It seems that Victor's later measurement has an effect on the earlier state of particles 1 and 4. This is would seem to saddle the realist with a commitment to backward causation, which many would find beyond the pale. Indeed, the authors themselves seem to take the experiment to show the inadequacy of the realist approach.

> If one views the quantum state as a real physical object, one could get the seemingly paradoxical situation that future actions appear as having an influence on past and already irrevocably recorded

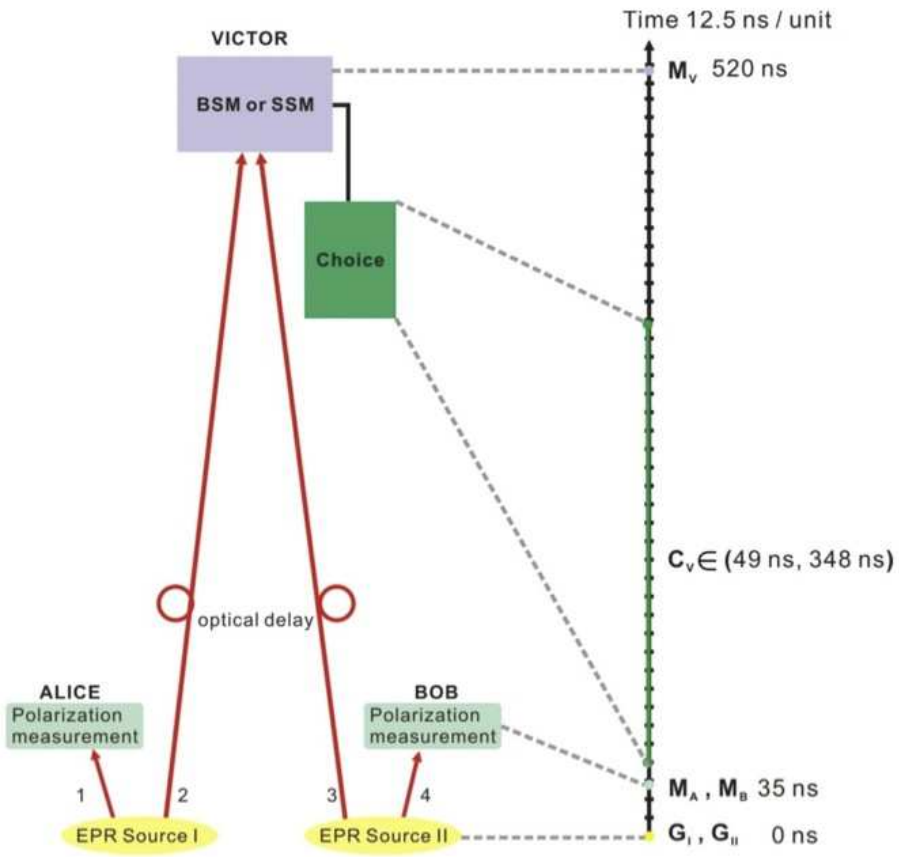


Figure 2: Delayed-choice entanglement swapping arrangement of Ma et al. (2012)

> events. However, there is never a paradox if the quantum state is viewed as to be no more than a "catalogue of our knowledge." (Ma et al. 2012, 483)

The committed realist must either deny that entanglement can be swapped from (2,3) to (1,4) in this case, or else provide some account of how it can occur. If one seeks to give the same explanation as in the case of entanglement swapping without delayed-choice, then they must allow that entanglement can obtain between (1,2) and (3,4) at the time of Victor's measurement. Of course, 1 and 4 do not exist at the time of Victor's measurement, so the entanglement relation must obtain between events at different times. We will return to this

idea below, but first, it's worth considering a way the realist may avoid this consequence.

4.1 Avoidance maneuvers

Matthias Egg (2013) offers a reply on behalf of the entanglement realist. He urges that to describe the foregoing as a genuine case of entanglement swapping is to beg the question against the realist. According to Egg's realist, the particles (1,4) are either entangled or not at the time of their detection (M_A, M_B), and later measurements cannot change this fact. In the case of entanglement swapping without delayed-choice, entanglement has been "swapped" to (1,4) as their quantum state has changed to become non-separable as a result of the measurement taken at Victor. Yet, according to Egg, the quantum state of (1,4) was separable when measured (M_A, M_B) in Ma's experiment and hence there was never a physical entanglement relation between (1,4) regardless of which later measurement Victor performs on (2,3).

So what should we make of the experimental evidence in favor of entanglement swapping to (1,4) after their detection?

> The Bell measurement on the [2,3] pair allows us to sort the [1,4] pairs into four subensembles corresponding to the four Bell states. Without delayed choice, this has physical significance, because each [1,4] pair *really is* in such a state after the [2,3] measurement. But if the [1,4] measurements precede the [2,3] measurement, the [1,4] pair *never is in any of these states*. This is entirely compatible with the fact that evaluating the [1,4] measurements *within* a certain subensemble shows Bell-type correlations. (Egg 2013, 1133)

Egg's reply focuses on an aspect of Ma's experiment that was omitted from the initial presentation. Unlike a simple EPR-B experiment, the correlations in the data recorded by Alice and Bob are only apparent once that data has been *sorted* into subsets ("subsenembles") according to the measurement performed and results obtained by Victor. Once we sort the results obtained by Alice and Bob in this way, we find that the subsets of data associated with Victor performing a BSM exhibit correlations indicative of entanglement.

Egg's point is that these correlations only appear once we sort the results in this manner, and such sorting needn't have any physical significance. It's unsurprising that correlations of *some* kind can be found when we conditionalize on the results obtained by Victor; after all, the photons measured by Victor were entangled with photons 1 and 4 until the latter were detected. Only when Victor's measurement actually causes a change in particles 1 and 4 are we justified in taking this process of sorting to have physical significance. Here we, as realists, should not allow that Victor's measurement has an effect on particles 1 and 4—doing so would require us to countenance backward causation—and hence the correlations obtained after sorting should not be taken to provide evidence for a genuine entanglement relation between particles 1 and 4.

4.1.1 Conflict with special relativity

Egg's reply requires that the entanglement realist make an important distinction between cases in which Victor's measurement occurs before Alice and Bob's measurements and those in which the time-order is reversed. Only the former, says Egg, are cases in which (1,4) are genuinely entangled. Yet, special relativity teaches that time-order is not an objective, frame-independent notion. If, for example, Victor's measurement (M_V) were spacelike separated from Alice and Bob's (M_A, M_B), then there would be no (frame-independent) fact of the matter about the time-order of the events. This scenario is not a mere hypothetical possibility either. In the much-publicized recent experiment of Hensen et al. (2015), entangled photon pairs are created via entanglement swapping from a location C that is spacelike separated from the measurement locations A and B (see Hensen et al. 2015, fig. 1e and 2a). Given such cases exist, adopting Egg's response would commit the realist to the claim that there is no (frame-independent) fact of the matter about whether the entanglement relation obtains. This would saddle the realist with a problematic sort of metaphysical indeterminacy.

In a footnote earlier in the paper, Egg offers the following rejoinder:

> Some of the most widely discussed realistic versions of quantum theory (e.g., Bohmian mechanics and the matter-density version of

> GRW) involve a commitment to a preferred foliation of spacetime. If these proposals are reasonable, then so is the assumption that there is a definite (although undetectable) temporal ordering between any two events. (2013, 1130, n.7)

It is of course true that a preferred foliation of space-time would solve the problem, and, indeed, this has been invoked in the service of some interpretations of quantum mechanics, but no such foliation (or a determinate time-ordering of spacelike separated events) is provided by our best theory of spacetime.[5] In order to take this option, the entanglement realist would be forced to claim that special relativity must be amended or at least, supplemented. This is a significant cost.

4.1.2 Parity of reasoning

Even if we ignore the conflict with relativity, there is a further worry with Egg's proposal.

The realist who would deny the reality of entanglement between (1,4) in the delayed-choice setup must claim that the standard argument for entanglement realism fails in this case purely because doing so leads to the undesirable result of backward causation. The argument for attributing entanglement in the ordinary swapping case relies only on the four-photon state (1) and the result obtained by Victor, without any mention of time. That same argument applied to the delayed-choice case delivers the same result, namely, that 1 and 4 are entangled. This result is confirmed by analyzing the data obtained by Victor, Alice, and Bob. Thus, there seems to be a tension in entanglement realism (so construed): on the one hand, it recommends recognizing a physical entanglement relation when it is instrumentally successful to do so, but, on the other hand, we should not posit such a relation in this case despite meeting the *very same* conditions that typically merit such an attribution. The failure to

[5] Of course, *general* relativity is our best theory of spacetime, and the situation there is more complicated. There are several candidates for a preferred foliation in general relativity, such as the "cosmological time" of relativistic cosmology. However, it is far from clear that any of these candidates should be taken to provide *the* metaphysically privileged way of carving up spacetime.

recognize entanglement in this case is an ad hoc measure to avoid the perceived alternatives of antirealism or backwards causation.[6]

5 Entanglement across time

If we reject Egg's attempt to reinterpret the outcome of these experiments, we are forced to consider whether the entanglement realism is consistent with genuine delayed-choice entanglement swapping. In particular, can the account of ordinary entanglement swapping be extended to cover the delayed-choice case? Because swapping is facilitated by entanglement relations, the answer to this question will depend on one's preferred metaphysics of entanglement. Suppose we adopt the action at a distance view. This would seem to saddle the realist with backward causal influence form Victor's measurement of (2,3) to particles 1 and 4 prior to their measurements by Alice and Bob.

But not so fast! First, we might question whether the influence is really *backward* in time. It is tempting to assume that Victor's measurement must bring about the earlier entanglement of 1 and 4, but the dependence between these events has a certain symmetry. Just as in the ordinary EPR-B case, it's hard to know which direction we should take the causal influence to go. Perhaps we should regard the earlier entanglement of (1,4) to cause the later BSM of (2,3). This might create worries about Victor's free will (or the randomness of the quantum random number generator), but these may not be decisive (see Evans et al. 2012, §7.1). Second, we might wonder whether entanglement-mediate influence should be understood as causal. It differs from paradigm instances of causation in many respects, including: (a) it fails to diminish with distance; (b) it cannot be shielded; (c) it doesn't involve a transfer of energy and; (d) it cannot be used to send signals. The last two conditions are of special importance as most paradoxes associated with backward causation seem to require them. In addition to these differences, action at distance must allow

[6]The sort of "instrumental success" I have in mind here is primarily the successful prediction of correlations in measurement results. We may also recall that in cases of entanglement swapping without delayed-choice, the attribution of an entangled state has important applications in quantum information theory. It is not unreasonable to suppose that related applications might be found for the attribution of an entangled state in the delayed-choice case as well.

for instantaneous influence to account for standard EPR-B experiments and, as a result, requires the rejection of the ordinary temporal asymmetry of cause and effect.

So, the action at a distance view can be extended to the timelike case without being committed to "backwards causation" by denying that either of the terms apply. Alternatively, one may countenance limited backward causation, but seek to downplay its significance for the reasons above (especially, the inability to use it for signaling).

Adapting the holist approach to allow for timelike entanglement is less straightforward. Part of the difficulty is due to the lack of clarity in the view generally. Many philosophers have advocated understanding entanglement in terms of a non-supervenient relation (e.g., Teller 1986; Howard 1985, 1989; Esfeld 2004); the entangled state of the joint system merits the attribution of a relation between its subsystems that fails to supervene on their individual intrinsic properties. This is sometimes paired with a claim that the compound system is more *real* or *fundamental* than the subsystems it comprises. One version may regard the joint system as a single object spread, smeared, or scattered across space. Another might take the distinct locations inhabited by the object to be unified in a more fundamental space of higher dimensionality.

The former case, in which joint systems are thought of as wholes scattered in space, seems to allow for extension to timelike separation without major problems. Temporally-scattered objects are not hard to imagine—a play with an intermission exists in two discontinuous timelike separated regions of spacetime—but, it's not obvious how such an approach is capable (on its own) of accounting for Bell-type non-local correlations. Indeed, Henson (2013) shows that the non-locality resulting from Bell's theorem is not avoided by denying separability. In some ways, this result is unsurprising. Merely redescribing the two photons in an EPR-B scenario as parts of a non-separable 2-photon whole does little to explain the correlations revealed by their measurement. This is not to say such an approach is hopeless, but it's unclear how it avoids the necessity of non-local influence.[7]

[7]It's possible that the advocate of this version of holism may wish to endorse action at a distance as well. Perhaps the reason *why* non-local influence is possible is that entangled systems form a

The other version of holism, in which joint systems are located at a single location in some higher-dimensional reality, promises to offer a more satisfying account of Bell-type correlations. The rough idea is to grant that the world is non-local in four-dimensional spacetime, but regard this as a reflection of a more fundamental space of higher-dimensionality which is entirely local (see Ismael 2012).

Yet, even if the higher dimensionality approach offers a promising alternative to action at a distance, it's not easy to see how the picture would be adapted to the case of timelike entanglement. The best known higher dimensionality view, wavefunction realism (Albert 1996; Lewis 2004; Ney 2013; Ney and Albert 2013), posits a fundamental ontology that includes the quantum wavefunction in a very high-dimensional configuration space. While such view may have the desired effect of eliminating spatial non-locality, time is left untouched.[8] The wavefunction *evolves* in configuration space with time. Thus, non-local influence among timelike separated regions would remain.

Could it be possible that timelike separated systems are reduced to a single object in a higher-dimensional space? Certainly. But, there are no known candidates for such a view. While there is talk of the emergence of space-time in some theories of quantum gravity, these ideas remain highly speculative. Furthermore, there is no reason to think that such theories will have the right features to provide a satisfactory account of entangled systems, much less those that are timelike separated.

6 Lessons for the metaphysics of entanglement

There are several lessons to be drawn. Most importantly, entanglement swapping doesn't undermine realism, but rather provides important insight into the

non-separable whole.

[8]It's unclear that wavefunction realism is able to account for entanglement in the manner suggested by Ismael. If *everything* is reduced to the wavefunction in high-dimensional configuration space, it doesn't seem able to account for what makes entangled systems special (c.f., Ismael and Schaffer 2013, 15).

nature of the entanglement relation. In particular, it compels the realist to revise certain aspects of their understanding of entanglement:

- Contrary to many presentations of the topic, entanglement does not require common preparation or previous interaction between entangled subsystems.
- Entanglement can account for changes in not just intrinsic (monadic) properties, but also the relations of entangled subsystems. Indeed, entanglement relations can beget new entanglement relations.
- Delayed-choice entanglement swapping can be accounted for in at least two ways:
  1. Following Egg, the realist can deny that genuine swapping occurs in delayed-choice setups.
  2. The realist can endorse the possibility of timelike entanglement.

  By taking the first option, the realist highlights their commitment to a time-ordering of spacelike events. Taking the latter option requires modifying the action at a distance or ontological holist views along the lines explored in the previous section.

I conclude by noting two very different potential sources support for timelike entanglement: (a) massless quantum fields in the Minkowski vacuum state (Olson and Ralph 2011, 2012) and (b) temporal analogues of Bell's theorem (Brukner et al. 2004; Fritz 2010). The import of these issues for a realist understanding of timelike entanglement remains to be seen.

References

Albert, D. Z. 1996. Elementary quantum metaphysics. In *Bohmian mechanics and quantum theory: An appraisal*, 277–284. Springer.

Berkovitz, J. 2016. Action at a distance in quantum mechanics. In *The Stanford Encyclopedia of Philosophy* (Spring 2016 ed.)., ed. E. N. Zalta.

Bohm, D. 1951. *Quantum Theory*. Dover Books on Physics. Dover Publications.

Bouwmeester, D., A. K. Ekert, A. Zeilinger, et al. 2000. *The Physics of Quantum Information*, Volume 38. Springer Berlin.

Brukner, C., S. Taylor, S. Cheung, and V. Vedral. 2004. Quantum entanglement in time. *arXiv preprint quant-ph/0402127*.

Egg, M. 2013. Delayed-choice experiments and the metaphysics of entanglement. *Foundations of Physics 43*(9): 1124–1135.

Einstein, A., B. Podolsky, and N. Rosen. 1935. Can quantum-mechanical description of physical reality be considered complete? *Physical review 47*(10): 777.

Esfeld, M. 2004. Quantum entanglement and a metaphysics of relations. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 35*(4): 601–617.

Evans, P. W., H. Price, and K. B. Wharton. 2012. New slant on the epr-bell experiment. *The British Journal for the Philosophy of Science*: axr052.

Fritz, T. 2010. Quantum correlations in the temporal clauser–horne–shimony–holt (chsh) scenario. *New Journal of Physics 12*(8): 083055.

Ghirardi, G., L. Marinatto, and T. Weber. 2002. Entanglement and properties of composite quantum systems: A conceptual and mathematical analysis. *Journal of Statistical Physics 108*(1-2): 49–122.

Healey, R. 2012. Quantum theory: A pragmatist approach. *The British Journal for the Philosophy of Science 63*(4): 729–771.

Hensen, B., H. Bernien, A. E. Dreau, A. Reiserer, N. Kalb, M. S. Blok, J. Ruitenberg, R. F. L. Vermeulen, R. N. Schouten, C. Abellan, W. Amaya, V. Pruneri, M. W. Mitchell, M. Markham, D. J. Twitchen, D. Elkouss, S. Wehner, T. H. Taminiau, and R. Hanson. 2015, 10). Loophole-free bell inequality violation using electron spins separated by 1.3 kilometres. *Nature 526*(7575): 682–686.

Henson, J. 2013. Non-separability does not relieve the problem of bell's theorem. *Foundations of Physics 43*(8): 1008–1038.

Howard, D. 1985. Einstein on locality and separability. *Studies in History and Philosophy of Science Part A 16*(3): 171–201.

Howard, D. 1989. Holism, separability, and the metaphysical implications of the bell experiments.

Ismael, J. 2012. What entanglement might be telling us. Unpublished draft, available online at `http://www.jenanni.com/papers/quantumholism-1.pdf`.

Ismael, J. and J. Schaffer. 2013. Quantum holism: Nonseparability as common ground. Unpublished draft, available online at `http://www.jenanni.com/papers/quantumholism-1.pdf`.

Ladyman, J., Ø. Linnebo, and T. Bigaj. 2013. Entanglement and non-factorizability. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 44*(3): 215–221.

Ladyman, J. and D. Ross. 2007. *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Lewis, P. J. 2004. Life in configuration space. *The British journal for the philosophy of science 55*(4): 713–729.

Ma, X.-s., S. Zotter, J. Kofler, R. Ursin, T. Jennewein, Č. Brukner, and A. Zeilinger. 2012. Experimental delayed-choice entanglement swapping. *Nature Physics 8*(6): 479–484.

Ney, A. 2013. Ontological reduction and the wave function ontology. *The Wave Function: Essays on the Metaphysics of Quantum Mechanics*: 168.

Ney, A. and D. Z. Albert. 2013. *The wave function: Essays on the metaphysics of quantum mechanics*. Oxford University Press.

Olson, S. J. and T. C. Ralph. 2011. Entanglement between the future and the past in the quantum vacuum. *Physical Review Letters 106*(11): 110404.

Olson, S. J. and T. C. Ralph. 2012. Extraction of timelike entanglement from the quantum vacuum. *Physical Review A 85*(1): 012306.

Peres, A. 2000. Delayed choice for entanglement swapping. *Journal of Modern Optics 47*(2-3): 139–143.

Psillos, S. 1999. *Scientific Realism: How Science Tracks Truth*. New York: Routledge.

Schaffer, J. 2010. Monism: The priority of the whole. *Philosophical Review 119*(1): 31–76.

Seevinck, M. 2006. The quantum world is not built up from correlations. *Foundations of Physics 36*(10): 1573–1586.

Teller, P. 1986. Relational holism and quantum mechanics. *British Journal for the Philosophy of Science*: 71–81.

Timpson, C. G. and H. R. Brown. 2010. Building with quantum correlations. *Quantum Information Processing 9*(2): 307–320.

Abstraction, Multiple Realizability, and the Explanatory Value of Omitting Irrelevant Details

Matthew C. Haug
The College of William & Mary
mchaug@wm.edu

Abstract

Anti-reductionists hold that special science explanations of some phenomena are objectively better than physical explanations of those phenomena. Prominent defenses of this claim appeal to the multiple realizability of special science properties. I argue that special science explanations can be shown to be better, in one respect, than physical explanations in a way that does not depend on multiple realizability. Namely, I discuss a way in which a special science explanation may be more abstract than a competing physical explanation, even if it is not multiply realizable, and I argue that this kind of abstraction can be used to support the idea that that special science explanation omits explanatorily irrelevant detail.

1. Introduction

Almost twenty years ago, Ned Block wrote of a long-standing "anti-reductionist consensus" in the philosophy of mind (and the philosophy of the special sciences more generally), which holds "that reductionism is a mistake and that there are autonomous special sciences" (1997, 107). Although this consensus seems to have been weakened in recent years, anti-reductionism is still arguably the dominant view. Disagreement remains about exactly what it is for a special science to be autonomous, but I take it that autonomy involves at least the following claim:

> *Explanatory autonomy*: Some special science explanation of a given fact or event is objectively better than any fundamental physical explanation of that fact or event.[1]

The classic papers that defended *explanatory autonomy*, and instituted the anti-reductionist consensus of which it is a part (e.g., Putnam (1967, 1975), Fodor (1974), and Kitcher (1984)), appealed to the alleged *multiple realizability* of special science properties. That is, these papers argued that any given special science

[1] *Explanatory autonomy* pits the explanations offered by fundamental physical theory against all other scientific explanations. In practice, however, debates about reduction have often been waged between explanations offered by a pair of "nearby" sciences, such as classical genetics and molecular genetics (e.g., Kitcher 1984) or cognitive psychology and neuroscience (e.g., Bechtel and Mundale 1999). Two common assumptions support this practice: first, that sciences are organized into hierarchical levels and, second, that sciences at the "lower levels" of this hierarchy (such as molecular genetics and neuroscience) are reducible to fundamental physics. Anti-reductionists and reductionists, then, disagree about whether sciences at higher levels in the hierarchy are reducible to relatively lower level sciences, with anti-reductionists claiming that reducibility fails to hold at some point in the hierarchy. Although there are reasons to doubt whether these assumptions are true, I shall adopt them in this paper for the sake of argument, along with the claim that neuroscience is reducible to fundamental physics. If these assumptions are false, then either anti-reductionism is in an even better position or the debate about reduction itself needs to be re-framed.

property could be realized by different (physically heterogeneous) physical properties and that this multiple realizability was crucial for defending *explanatory autonomy*.[2] In this paper, I argue that there is at least one respect in which some special science explanations are better than physical explanations that need not be "based on an argument from multiple realizability." Namely, some special science explanations isolate explanatorily relevant features even if they do not involve properties that are multiply realizable. As I discuss below, these explanations are more *abstract* than competing physical explanations in a way that does not entail that they are more general. If I am right, it is a mistake to think that *explanatory autonomy* hinges on multiple realizability. Avoiding this mistake sidesteps the morass of debates about the nature and extent of multiple realizability and opens up new conceptual space in an old debate.

2. Relevance, Generality, and Multiple Realizability

One plausible way in which an explanation is objectively good is if it provides the right amount of explanatory detail (see, e.g., Garfinkel 1981; Batterman 2002; Strevens 2008). For instance, suppose that an extremely loud utterance of the word "shatter" causes a crystal wine glass to break (cf. Dretske 1988, 79). An explanation that cites the pitch and intensity of the sound waves is better than one that also cites the meaning of the utterance. This is because the meaning of the utterance was irrelevant to the glass's breaking. The glass would have broken regardless of the utterance's meaning (or lack thereof). Examples like this one support the following explanatory virtue:

> *Relevance*: Other things being equal, one explanation is better than another if it includes fewer features that are irrelevant to (the production of) the fact or event to be explained.

One explanation will be better, with respect to *relevance*, than another to the extent that it is more *abstract*, in the sense that it eliminates or omits details.

There are other good-making features of explanations aside from *relevance*. For instance, good explanations will unify disparate phenomena. Other things being equal, one explanation is better than

[2] As Block notes, it is part of the consensus that the autonomy of the special sciences is "based on an argument from multiple realizability" (1997, 107).

another if it is more *general*, if it is, in some sense, more widely applicable. One kind of generality that has often been used to defend *explanatory autonomy* traces back to Putnam's (1967, 1975) work: one explanation of a given fact or event is more general, in this sense, than another if applies to wider range of (logically) possible situations. As with *relevance*, explanations will be more general to the extent that the properties that figure in them are more abstract.[3] By eliminating or omitting physical details, some abstract special science properties will apply to more possible situations.

Since both the greater relevance and greater generality of special science explanations depend on their being more abstract than physical explanations, it can seem as if these two explanatory virtues stand or fall together. If there were a single kind of abstraction that was required to support both of these virtues, then this would go a long way toward explaining why influential defenses of the greater *relevance* of special science explanations (and not just their greater *generality*) have relied on multiple realizability. Here I briefly sketch two such defenses.

First, consider Putnam's famous discussion of why a 15/16" rigid cubical peg fits through a 1" square hole, but not a 1" circular hole, in a rigid board. Putnam claims that an explanation of these facts in terms of the shapes and relative sizes of the peg and holes and the rigidity of the peg and board "brings out" the "*relevant structural features of the situation*," which any microphysical explanation "conceals" (1975, 296, 297, italics in original). And he suggests that this is because the "higher level" structural features are multiply realizable: the same explanation in terms of these features will be correct "whether the peg consists of molecules, or continuous rigid substance, or whatever" (ibid., 296). The idea is that changing the underlying microstructure makes no difference to whether the square peg goes through the round hole; all that is relevant is the abstract structural features that are *common* to those different physical realizers.

Similarly, Kitcher (1984) claims that a derivation of the general principles of classical genetics from molecular genetics is not explanatory because "in charting the details of the molecular

[3] In fact, in a paper that isolates and defends this kind of generality as a dimension of explanatory depth, Brad Weslake (2010) simply calls it *abstraction*.

rearrangements the derivation would only blur the outline of a simple cytological story, *adding a welter of irrelevant detail*" (ibid., 347, italics added). A little later, he claims that "adding [molecular] details would only disguise the relevant factor," namely, that meiosis is a "pair-separation process" (ibid., 348). And, he claims that this fact about relevance hinges on multiple realizability—on the alleged fact that pair-separation processes are "heterogeneous from the molecular point of view" (1984, 349) because they are "realized in a motley of molecular ways" (ibid., 350).[4]

Basing the greater *relevance* of special science explanations on multiple realizability may be ill advised, however. For, almost contemporaneously with the formation of the anti-reductionist consensus, some philosophers raised serious doubts about the multiple realizability argument,[5] and, in the two decades since Block's (1997) article appeared (in which he responded to one of these lines of doubt) a number of authors have raised additional critiques concerning the extent of multiple realizability and whether it can be used to establish the explanatory superiority of the special sciences.[6] For all I say in this paper, the multiple realizability argument may be successfully defended against such critiques. However, if I am right, such a defense is unnecessary: the greater *relevance* of some special science explanations can be supported by a kind of abstraction that need not involve multiple realizability.

3. A Distinction Among Kinds of Abstraction

One entity is more abstract than another if it lacks detail that the other possesses. Different instances of abstraction will involve omission or elimination of different kinds of details, some of which I discuss below. There is an important distinction to be drawn among these varieties of abstraction: some kinds of abstraction require that more abstract entities are more *general* than less abstract ones (call this *generality-entailing (GE) abstraction*), while other kinds of abstraction do not impose this requirement

[4] See also Garfinkel (1981), who claims that a "microexplanation" of why a rabbit was eaten (in terms of the "equations of interaction between individual foxes and individual rabbits (depending on such things as their physiology and reaction times)" (ibid., 55)) is inferior to a "macroexplanation" that simply cites the high population-density of foxes in the area (and implicitly, the Lotka-Volterra equation). Garfinkel claims that this is because the microexplanation "contains much that is *irrelevant* to why the rabbit got eaten and … [these irrelevant data] *bury* the explanation unrecognizably" (ibid., 56, italics added**).** And *this* is true, according to Garfinkel, because the macroexplanation is *stable* through variation in the way that the high population-density of foxes is realized (ibid., 57).

[5] See, e.g., Lewis (1969), Kim (1972, 188-92).

[6] See, e.g., Bechtel and Mundale (1999); Shapiro (2004)

(call this *non-generality-entailing (NGE) abstraction*). I first discuss two varieties of abstraction that fall on the former side of this divide; then, I outline two kinds of abstraction that fall on the latter side.

First, consider the way in which determinables are more abstract than their determinates. *Being red* is more abstract than *being crimson* or *being scarlet*. What kind of detail does a determinable lack that its determinates possess? Assuming that colors are individuated by three "determination dimensions": hue, saturation, and brightness, *being crimson* will be associated with particular values (or a small range of values) of hue, saturation, and brightness. It will occupy a relatively small portion of a three-dimensional color space. *Being red*, by contrast, will not be specified at this fine-grained level of detail. Rather, it will characterized by a broader range of hue, saturation, and brightness values. It will occupy a larger portion of three-dimensional color space, a portion that includes the space occupied by *being crimson* as a proper subspace. Following Haug (2011), I'll call this kind of abstraction, *homotopic* abstraction, since it applies to properties that are characterized by the *same* property *space*. In general, we can say that a property P is more homotopically abstract than a property Q if and only if Q occupies a proper subspace of the portion of the property space occupied by P.[7]

Second, consider the relation between a multiply realizable property and each of its individual realizers. There is disagreement about how to characterize multiple realizability, but on any plausible account, multiply realizable properties are more abstract than their individual realizers. If property P (of type X) is multiple realizable, then P omits (or "abstracts away") from features that are unique to its individual realizers and isolates features that these realizers (or the objects that possess them) have in common. This will be true whether these features are causal powers (Wilson 1999; Shoemaker 2001; Gillett 2002), or functional roles (Shapiro 2004; Polger 2007), or exact similarity of X-type features (Funkhouser 2007).

[7] Note that this differs somewhat from the definition given by Haug (2011). Eric Funkhouser (2006, 2014) provides a helpful discussion of "determination dimensions" and how they (together with "non-determinable necessities") specify the nature of kinds of properties. Note that homotopic abstraction is the inverse relation of what Funkhouser calls "specification."

Homotopic abstraction and multiple realizability both clearly require more abstract properties to be more general than less abstract ones: they are instances of GE-abstraction. Clearly, the property *being red* must apply not just to scarlet objects but also to crimson ones. Similarly, a multiple realizable property, P, must apply not only to objects that possess one of P's realizers, R, but also to objects that possess another of P's realizers (and do not have R).

Now consider a third example of abstraction. Given a physical system, one model or set of equations describing that system will be more abstract than another if it eliminates or omits at least one feature of that system (e.g., a quantity, degree of freedom, or boundary or initial condition) that the latter model or set of equations includes. For example, suppose we have a rotating sphere moving with a constant linear velocity along the x-axis. A set of equations (*p*) that includes only an equation of motion for the sphere's linear motion is more abstract than one (*pL*) that includes equations of motion for *both* its linear motion *and its rotation*. Similarly, a thermodynamic model of a gas that takes the number of particles and volume to be infinite (while maintaining a constant ratio of number of particles to volume) is more abstract than a model that includes the boundary conditions imposed by the gas's container. (For these examples, see Knox (2016, 44-45, 50).)

Finally, consider the relation between *having some hue or other* (i.e., *being hued*) and *having some color or other* (i.e., *being colored*). *Being hued* is more abstract than *being colored*; it omits the other two dimensions of color: saturation and brightness. Following Haug (2011), we can call this kind of abstraction, *heterotopic abstraction*, since it applies to properties from *different* property *spaces*. If a property Q metaphysically necessitates a property P, then P is more heterotopically abstract than Q if and only if the characteristic property space of P has fewer dimensions than the characteristic property space of Q.[8]

These two kinds of abstraction do *not* require that more abstract entities are more general than less abstract ones. A more abstract model or set of equations need not apply to more systems than a less abstract model or set of equations. For example, suppose that we take the set of equations *pL* above and

[8] Note that this also differs slightly from the account given by Haug (2011).

stipulate that the linear momentum is zero; call this set of equations (including the trivial equation of motion for rotation) *pL-zero*. Then, the set of equations *p* (that includes only an equation for linear motion) will apply (i.e. accurately describe) the kinematics of *exactly the same* set of possible physical systems as *pL-zero*, namely, those spheres moving along the x-axis that have zero angular momentum. However, *p* is still more abstract than *pL-zero* since it *omits* the equation of motion for rotation entirely.[9]

Similarly, unlike the pair of *being red* and *being scarlet*, *being hued* is not more general than *being colored*. Necessarily, anything that has a hue is also colored (and, necessarily, anything that is colored is also hued). These properties are necessarily co-extensive, even though one is more abstract than another. Thus, in these cases, greater abstraction does not require greater generality: they are instances of NGE-abstraction.

4. NGE-abstraction and Explanatory Relevance

With the distinction between GE-abstraction and NGE-abstraction in hand, we can see that the former (and thus multiple realizability) is not required for defending the greater *relevance* of special science explanations; NGE-abstraction will work at least as well.

First, suppose that we have non-rotating sphere moving along the x-axis with a constant velocity. The systems of equations *pL-zero* and *p*, from above, apply to exactly the same systems; *p* is not more general than *pL-zero*; nevertheless, *p* clearly provides a better explanation of the sphere's position; it isolates only the features that are relevant to this explanandum, omitting the irrelevant (and, incidentally, zero-valued) angular momentum. (For a more substantive example of NGE-abstraction with respect to certain features of a system, see Section 5.)

Now, return to Putnam's example from Section 2, and note that the "high level" structural properties in Putnam's case are more heterotopically abstract than the microphysical properties that underlie them. *Being a circle* and *being a square* each have a single determination dimension (diameter length and side length, respectively). Further, *rigidity* (also known as *stiffness*) also has a single determination

[9] This example is inspired by Knox's (2016, 52) discussion of idealization as a "precursor to abstraction."

dimension measured in units of [force/distance]. By contrast, the microphysical properties that are possible realizers of these structural features are characterized by more determination dimensions, including at least: (1) the nature of the components that make up the peg and board, (2) the nature of the bonds that hold between these components (e.g., ionic, covalent, metallic, or none (in the case of a (logically possible) continuous rigid substance)); (3) the angle(s) between these bonds; and (4) the entropy of the peg and board (which is especially important for understanding the stiffness of polymers such as rubber). (See Roylance (2000) for a discussion of the microscopic basis of *stiffness*.)

A correct explanation of the fact that the square peg does not pass through the round hole depends only on the relations between the determination dimensions of *being a circle*, *being a square*, and *rigidity*. The determination dimensions of the underlying realizers are irrelevant to this explanation. Importantly, it is not the *multiplicity* of the *values* of these other determination dimensions that is crucial here but simply the fact that these other dimensions characterize the microphysical realizers *at all*. That is, it is not facts about homotopic abstraction or multiple realizability that are crucial for *relevance* but rather facts about heterotopic abstraction, a kind of *non-generality-entailing* abstraction.[10]

One might worry that multiple realizability (or homotopic abstraction) is required to justify the claim that these other dimensions are in fact explanatorily *irrelevant* to the explanandum—that without showing that the explanatory relationship is "stable" or "robust" through *variation* in the *values* of these underlying dimensions we would have no evidence that these other dimensions are *themselves* explanatorily irrelevant. However, this is not the case. Even without showing that the generalization that 15/16" square pegs do not fit through 1" circular holes is independent of variation in the underlying physical realizers, it remains the case that this generalization is realization independent in the sense that one can discover and confirm it without knowing anything about the physical realizers that underlie it (much less confirming a generalization between them). That is, the very fact that structural predicates like

[10] Haug (2011) claims that heterotopic abstraction can support the idea that some special science explanations omit explanatorily irrelevant details that are included in physical explanations. However, the discussion of this point is very brief, and Haug does not discuss how other forms of NGE-abstraction can also support this idea. (Cf. the discussion of the sphere example above and of thermodynamics in Section 5.)

"is a 15/16" square peg" are projectible is strong presumptive evidence that such predicates pick out objective features of the world that figure in real regularities. (On this point, see Antony (1999, 14ff.).) The underlying microstructure (whether it maps many-to-one *or one-to-one* to macrostructure) is irrelevant to confirming that such regularities obtain.

We can get clearer about how heterotopic abstraction can be used to defend the greater *relevance* of special science explanations without appealing to multiple realizability by seeing how such a defense effectively blocks reductionist appeals to "disjunctive properties" or "local reductions" (see, e.g., Kim 1992; Lewis 1994). These appeals, in effect, claim that (perhaps disjunctive) physical properties can in fact be matched up one-to-one with (at least structure- or species-restricted) special science properties and that explanations in terms of these physical properties will be just as *relevant* and *general* as (any genuine) special science explanations. However, while these reductionist gambits may be successful with respect to *generality*, facts about heterotopic abstraction show that they are ineffective at undermining the greater *relevance* of special science explanations.

Consider a disjunctive property, D, that has every metaphysically possible realizer of *being a 15/16" square peg* as a disjunct. D is just as general as *being a 15/16" square peg*; they are necessarily co-extensive. Thus, Putnam's multiple realizability argument for the greater *relevance* of the macro-structural explanation is undermined.[11] However, D's instantiation consists merely in one of its disjuncts being instantiated, so its property space is arguably the *sum* (i.e. the span of the union) of the property spaces of each of its disjuncts. That is, D's set of determination dimensions is the union of the sets of determination dimensions of its disjuncts. Thus, *being a 15/16" square peg* is still more heterotopically abstract than D; it still isolates features that are explanatorily relevant to the fact that the peg fails to pass through the round hole, while omitting features that are irrelevant to this behavior.

[11] Implicit in the above discussion is Putnam's claim that "the higher level explanation is far more general [than the microphysical one], which is why it is *explanatory*" (1975, 297, italics in original). But this claim is not true when the microphysical explanation is in terms of D.

Funkhouser also responds to the "disjunctive property" objection to the multiple realizability argument by pointing out that a realized property R has different determination dimensions than a disjunctive property whose disjuncts are each of R's possible realizers (2014, 108-9). But Alexander Bird (2015) rightly asks how this point alone constitutes a reply to the objection. The key point that Funkhouser leaves out is that a difference in determination dimensions—in particular, greater heterotopic abstraction—supports greater explanatory *relevance* even if the absence of multiple realizability. Funkhouser emphasizes that realized properties and their realizers, by having different determination dimensions, are at different "levels of abstraction" (2014, 78, 89, 124). However, he does not seem to recognize fully that (as the disjunctive realizer case illustrates) this kind of abstraction does *not* depend on *multiple* realizability but rather on the realization relation (a particular kind of asymmetric necessitation) itself.

5. Implications for How to Frame Debates about Reduction and Autonomy

In the last few years, several philosophers have defended explanatory autonomy or irreducibility without relying on multiple realizability (e.g., Wilson 2010; Knox 2016). I'll conclude by briefly applying the notion of NGE-abstraction to one of these defenses (Knox 2016) and suggesting that this can help us better understand how to formulate debates about explanatory autonomy and reducibility, in general.[12]

Knox (2016) argues that thermodynamics offers novel explanations of many phenomena, such as why diesel engines, unlike gasoline engines, do not need spark plugs. The novelty of this thermodynamical explanation consists in the fact it involves an abstraction (namely, omitting all details related to heat transfer) that "cuts across" the quantities that are recognized as natural by statistical mechanics (ibid., 46, 56). Further, Knox claims that this fact about explanatory novelty is compatible with the *reducibility* of thermodynamics to statistical mechanics. According to Knox, the key to

[12] Wilson's (2010) argument that "eliminations of degrees of freedom" are sufficient for ontological irreducibility also relies on NGE-abstraction and not multiple realizability, but a discussion of this fact (and its implications) will have to await another occasion.

understanding this compatibility is that the bridge laws linking thermodynamical and statistical mechanical quantities will involve complex, "mathematically irreversible" operations (such as taking the limit as the number of gas particles goes to infinity) and thus will themselves involve abstraction (ibid., 54, 57). As a result, further abstractions with respect to thermodynamical quantities will be "opaque" from the perspective of statistical mechanics (ibid., 42).

First, note that these further abstractions will be instances of the first kind of NGE-abstraction discussed above—ignoring heat transfer simplifies the thermodynamical equations and isolates the explanatorily relevant features. Knox claims that this results in "novel explanations that are not merely abstractions of some more detailed [statistical mechanical] picture" (ibid., 41), and she seems to ground this novelty in the "mathematically irreversible" change in variables that occurs when one moves from statistical mechanical to thermodynamical quantities (ibid., 56). However, not every mathematically irreversible operation seems to induce this kind of "unnaturalness" from a lower level perspective: for example, taking a sum or a mean will lead to loss of information, but the result *will* be "merely an abstraction from some more detailed underlying picture." I think that the second kind of NGE-abstraction—heterotopic abstraction—can help here. It is not mere "mathematical irreversibility" that supports novelty but *heterotopic abstraction*. The mean of a quantity of a system is more homotopically abstract than a description that specifies the particular values of that quantity had by the components of that system: it omits detail within a *single* property space. But moving from a statistical mechanical quantity to a thermodynamical quantity involves moving to a *new* property space, one whose dimensions "cut across" the dimensions of the property space of the statistical mechanical quantity.

Knox notes that her account of explanatory novelty fits poorly into standard taxonomies of emergence/reduction that characterize weak emergence as merely epistemic and strong emergence as metaphysical (2016, 58). Her account is weaker than standard epistemic accounts in that it is compatible with the theoretical reduction of thermodynamics to statistical mechanics, but it is stronger than

epistemic accounts in that it depends on objective features of the world and not merely on our cognitive limitations (ibid., 58, 44).

But how can there be any *objectively* better "high level" explanations if all of the properties involved in those explanations can be mapped one-to-one via *bridge laws* to "low level" properties? What "objective features of the world" could these explanations be tracking other than those of fundamental physics? I think that NGE-abstraction suggests a framework within which to answer these questions. A property can be more NGE-abstract than another with which it is necessarily co-extensive. If NGE-abstraction is itself an "objective feature of the world," this suggests that we should adopt a *hyperintensional* criterion for property individuation.[13] In short, there are "objective features of the world" that are more fine-grained than, and thus cannot be captured by, the resources used in standard formulations of the metaphysics of reduction and autonomy. Working out this hyperintensional account of property individuation will not be a trivial task, but if the discussion in this paper is on the right track, it is a task that is important not only for defending the autonomy of the special sciences but also for spelling out exactly what such autonomy amounts to.

[13] This provides a further motivation for adopting a program of "hyperintensional metaphysics" (Nolan 2014), one that is more closely tied to actual scientific practice than recent work on the hyperintensional notion of "grounding" in analytic metaphysics.

References

Antony, Louise. (1999) "Multiple Realizability, Projectibility, and the Reality of Mental Properties." *Philosophical Topics*. 26: 1-24.

Batterman, Robert W. (2002) *The Devil in the Details*. Oxford: Oxford UP.

Bechtel, William and Jennifer Mundale. (1999) "Multiple Realizability Revisited: Linking Cognitive and Neural States." *Philosophy of Science*. 66: 175-207.

Bird, Alexander. (2015) "Review of *The Logical Structure of Kinds*." *Notre Dame Philosophical Reviews*. 23 Oct. https://ndpr.nd.edu/news/61948-the-logical-structure-of-kinds/

Block, Ned. (1997) "Anti-reductionism Slaps Back." *Philosophical Perspectives*. 11: 107-132.

Dretske, Fred. (1988) *Explaining Behavior: Reasons in a World of Causes*. Cambridge MA: MIT Press.

Fodor, Jerry. (1974) "Special Sciences." *Synthese*. 28: 97-115.

Funkhouser, Eric. (2006) "The Determinable-Determinate Relation." *Noûs*. 40: 548-569.

———. (2007) "A Liberal Conception of Multiple Realizability." *Philosophical Studies*. 132: 467-494.

———. (2014) *The Logical Structure of Kinds*. Oxford: Oxford UP.

Garfinkel, Alan. (1981) *Forms of Explanation*. New Haven, CT: Yale UP.

Gillett, Carl. (2002) "The Dimensions of Realization: A Critique of the Standard View." *Analysis*. 62: 316-323.

Haug, Matthew C. (2011) "Abstraction and Explanatory Relevance; or, Why Do the Special Sciences Exist?" *Philosophy of Science*. 78: 1143-1155.

Kim, Jaegwon. (1972) "Phenomenal Properties, Psychophysical Laws, and the Identity Theory." *The Monist*. 56: 177-192.

———. (1992) "Multiple Realizability and the Metaphysics of Reduction." *Philosophy and Phenomenological Research*. 52: 1-26.

Kitcher, Philip. (1984) "1953 and All That." *Philosophical Review*. 93: 335-373.

Knox, Eleanor. (2016) "Abstraction and Its Limits: Finding Space for Novel Explanation." *Noûs*. 50: 41-60.

Lewis, David. (1969) "Review of *Art, Mind, and Religion*." *Journal of Philosophy*. 66: 23-35.

———. (1994) "Reduction of Mind." in *Companion to the Philosophy of Mind*. Ed. Samuel Guttenplan. Cambridge: Blackwell. 412-431.

Nolan, Daniel. (2014) "Hyperintensional Metaphysics." *Philosophical Studies*. 171: 149-160.

Polger, Thomas. (2007) "Realization and the Metaphysics of Mind." *Australasian Journal of Philosophy*. 85: 233-259.

Putnam, Hilary. (1967) "Psychological Predicates." In *Art, Mind, and Religion*. Eds. W.H. Capitan and D.D. Merrill. Pittsburgh: University of Pittsburgh Press, 37-48. Reprinted as "The Nature of Mental States" in his *Mind, Language, and Reality*: *Philosophical Papers, Vol. 2*. Cambridge: Cambridge UP. 429-440.

———. (1975) "Philosophy and Our Mental Life." In his *Mind, Language, and Reality*: *Philosophical Papers, Vol. 2*. Cambridge: Cambridge UP. 291-303.

Roylance, David. (2000) "Atomistic Basis of Elasticity." *Mechanics of Materials*. (MIT OpenCourseWare) http://ocw.mit.edu/courses/materials-science-and-engineering/3-11-mechanics-of-materials-fall-1999/modules/elas_2.pdf

Shapiro, Lawrence. (2004) *The Mind Incarnate*. Cambridge, MA: MIT Press.

Shoemaker, Sydney. (2001) "Realization and Mental Causation." In *Physicalism and Its Discontents*. Ed. Carl Gillett and Barry M. Loewer. Cambridge: Cambridge UP. 74-98.

Strevens, Michael. (2008) *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard UP.

Weslake, Brad. (2010) "Explanatory Depth." *Philosophy of Science*. 77: 273-294.

Wilson, Jessica. (1999) "How Superduper Does a Physicalist Supervenience Need to Be?" *Philosophical Quarterly*. 49: 33-52.

———. (2010) "Non-Reductive Physicalism and Degrees of Freedom." *British Journal for the Philosophy of Science*. 61: 279-311.

Disambiguating Latent Variables

ABSTRACT

In contrast to Borsboom (2008) who distinguishes between manifest and latent variables on epistemic grounds in terms of "epistemic accessibility," I advocate a demarcation on pragmatic grounds. The latter way of understanding this distinction does justice to the intuitions driving Borsboom's account, but avoids unnecessary epistemic complications. I then turn to two cases, the Flynn Effect and the case of psycho-educational assessment, and show an equivocal understanding of one latent variable, Spearman's *g*, has led some researchers to draw paradoxical conclusions regarding cognitive ability.

Word count, including references and footnotes: 4790

Disambiguating Latent Variables

1. Introduction: Variables Latent, Variables Manifest

Latent variables are ubiquitous in the social and behavioral sciences. Some claim they are an indispensable part of social and psychological research (Sobel 1994). We may distinguish between variables that are *manifest* and variables that are *latent*. Manifest variables are, at first pass, distinguished by being observed; at least this is one popular way of distinguishing them from latent variables. Suppose we set out to measure the lengths of various objects. With a meter stick in hand, we get to work: the length of the swimming pool is 100 meters, the ceiling is three meters from the floor, etc. In these cases, length is a manifest variable. It is also an observed property of the objects who length we measure. Some methodologists claim that what distinguishes manifest from latent variables is whether the quantity in question is observed or is inferred from some observed measure, respectively. While being observed and being manifest are seemingly concomitant properties, I will argue that this concomitance is not philosophically significant, and, furthermore, by avoiding a criterion that demarcates latent and manifest variables in terms of whether a property instance is observed, certain quagmires can be sidestepped.

Contrast the case of manifest variables with the following: we want to find out someone's or a demographic group's socioeconomic status (SES). However, instead of setting out with an instrument that measures SES directly, we pass out questionnaires asking our subjects to report their gross annual income, level of education and occupations of both the subject and his parents. Based on the values for each of those variables, we locate our

subjects on the SES index. SES is not observed directly, it is a composite score based on values for manifest variables. Similarly for psychometric research on cognitive ability, a battery of tests is administered to an individual and the test-scores are manifest variables the correlations among which are explained by positing a latent variable, *g* (taken to denote general intelligence). Epistemic access to *g* is mediated by measuring its manifest indicators; no direct measure of general intelligence is available, though some indicators are taken to be better measures of it than others.

Variables considered to be manifest in one context would be considered latent in a different context if they figure in as a latent variable in a latent variable model. A latent variable model specifies the relationship between various observable indicators (manifest variables) and a latent variable or class of latent variables. Hence, height, though a manifest variable in the examples considered earlier, could be a latent variable depending on one's measurement methods. For example, we may want to assess the heights of adolescents in a town where there are no meter sticks, rulers, or other devices for measuring height directly. Instead we may record their weights and shoe sizes (as indicated by the label in the shoe) and use those values to infer values for height. On the basis of those measures we should be able to predict with good accuracy the length of the adolescents since the values for the manifest variables are known to correlate highly with height in adolescents; note that in this context there was no appeal to observation as a distinguishing characteristic.

What makes height a latent variable in this example is that it and its values for a given subject are inferred on the basis of known indicators of height *in this measurement scenario*. Hence, whether a variable is latent depends on the method used to ascertain its value in a

particular instance. The latent/manifest distinction does not track or commit one to the so-called “observable/unobservable” or “observational/theoretical” distinctions. This is all the better for latent variable modelers that the legitimacy of their method not piggyback on controversial distinctions.

Latent variables form a heterogeneous bunch, united only by the fact that they are not manifest. I will restrict my investigation of latent variables to the social and behavioral sciences and the inferential problems introduced by positing latent variables, namely how we go from latent variables to quantities in nature. Specifically I will be concerned with latent variables in psychometrics, a branch of psychology devoted to the investigation of psychological traits and the structure of individual and group differences in psychological traits. I will devote considerable attention to the general factor of intelligence, i.e., the *g*-factor and also consider latent variables in other disciplines such as socioeconomics.

Unsurprisingly there are alternative views on how to understand the distinction between latent and manifest variables. Borsboom (2008) argues that the distinction is epistemic; it maps onto the differential evidential gap between data and their causes. Borsboom uses the term ‘observed’ to mean ‘manifest’. On this account, in the case of manifest variables, we assign probability equal to one to the measurement outcome; in the case of latent variables, we assign probabilities less than one to measurement outcomes.

Furthermore, Borsboom’s criterion for demarcating variables is either too strong or arbitrary. His criterion of certainty is satisfied in very few measurement contexts, even those in which we would be inclined to deem the variable patently manifest. For example, the more times I concatenate rulers, the less confident I am that I have not made some measurement

error, even if for each time I concatenate I believe that I have done so without error. Using a ruler to measure the length of a standard sheet of paper (or another ruler) is one thing, but measuring moderately large distances is another. On Borsboom's account, length becomes latent once my confidence drops below one. But here charges of arbitrariness arise: why must confidence amount to probability equal to one for the variable to be an "observed" variable? Without justifying that threshold, it seems arbitrary to set it at certainty and unnecessarily stringent.

I suggest that we demarcate variables pragmatically: simply read off their status from the measurement or structural equation model. If in the model length is treated as a latent variable, then length is a latent variable *in that model* (e.g., in the structural equation modeling package LISREL, the variable is indicated by an ellipse instead of a rectangle, or such as a regression coefficient in a regression model). This approach is contextualist: a variable's status depends upon the measurement context. It seems that in many circumstances drawing the distinction in this way will make sense of Borsboom's idea that we seem to have better epistemic access to manifest variables, or perhaps *vice versa*: Borsboom's idea explains why we allow for some variables to be treated as latent in our models. The contexts in which one treats length as a latent variable are likely to be those in which there is an epistemic gap between what one is measuring "directly" and length. Likewise, if I can measure length itself I am unlikely to treat it as latent in my model. However, that a variable is latent is conceptually independent of my epistemic situation; it is contingent upon the formal aspects of the model. Drawing the line between latent and observed variables this way avoids charges of arbitrariness or immoderate stringency.

2. Interpretation and Latent Variables

Though latent variables may be invoked to refer to unobservable objects, such as electrons or quarks, as they factor in psychometrics and the social sciences, latent variables are typically taken to refer to properties, e.g., personality characteristics such as “extraversion” and abilities such as “general intelligence.” I will assume that properties, or at least property instances, have causal powers. Thus, if a latent variable successfully refers to a property or property instance, then the variable’s referent has causal powers (i.e., it is causally efficacious). This commitment rules out the possibility of epiphenomenal latent variables, and this might seem suspect given that I am dealing with latent variables that purportedly refer to mental properties. Psychometrics seems to presuppose that epiphenomenalism is false, since psychological attributes, the referents, of latent variables are alleged to be causally efficacious if they exist at all.[1]

[1] However, some psychometricians believe that a psychological attribute, A, is causally efficacious only if there is actual variability in A. That is, a disposition to effect change is not strong enough. This variability in position on A is manifested in test behavior. See Borsboom (2005) and Holland (1986). One interesting consequence of this position is that general intelligence, the purported referent of the *g*-factor, is not causally efficacious since it exhibits no variability within individuals, i.e., there is no intraindividual variability. The severity of this consideration for theories of intelligence that take general intelligence to be a central theoretical posit and whether interindividual variation is sufficient to save general intelligence *qua* causally efficacious psychological attribute are questions that merit further

One may advance a wholesale rejection of psychometric constructs as meaningless or mere statistical artifacts. However, my starting point is psychometric practice, for it is this practice that I wish to clarify. Rejecting the entire discipline would be not only a disservice to a scientific discipline which shows no sign of losing steam, but it would also be a disservice to the philosophy of science which potentially stands to gain from careful examinations into psychological measurement (see Trout 1999; Sesardic 2000).

2.1 Latent Variable Modeling as Data Reduction

Factor analysis is one statistical procedure for discovering latent variables (exploratory factor analysis) and confirming latent variable models (confirmatory factor analysis). The utility of factor analysis is manifold. First, factor analysis is a data reduction technique. Suppose you have a p×p correlation matrix. The correlated items may be performance on psychometric tests or what have you. The larger p is, the greater the number of correlations in the matrix and also the more unwieldy the matrix becomes. Sometimes it might be useful to express the information contained in the correlation matrix with a smaller number of variables. For example, it may be more economical and cognitively tractable to deal with a 5×20 factor matrix expressing the relationship between the manifest variables and a compendious set of latent factors, rather than a 20×20 correlation matrix. To illustrate analysis, consider the following 8×8 correlation matrix from Jensen (1998, 80):

	V1	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V8**
V2	.5600							

attention.

V3	.4800	.4200						
V4	.4032	.3528	.3024					
V5	.3456	.3024	.2592	.4200				
V6	.2880	.2520	.2160	.3500	.3000			
V7	.3024	.2646	.2268	.2352	.2016	.1680		
V8	.2520	.2205	.1890	.1960	.1680	.1400	.3000	
V9	.2016	.1764	.1512	.1568	.1344	.1120	.2400	.2000

Table 1: Hypothetical correlation matrix of intelligence test data

As the number of variables increases, so does the utility of being able to represent the information in terms of a few latent variables. Note that some of the indicators (the V's) correlate more strongly with each other than with others. For example, V1, V2, and V3 are more strongly mutually correlated than they are with other variables. The correlations between variables can be expressed more economically in terms of a correlation with a latent variable. Factor analysis enables us to transform the correlation matrix above into a *factor matrix* expressing the correlation between each test and an "underlying" factor (table 2).

	1st order			**2nd order**
Variable	**F1**	**F2**	**F3**	***g***
V1	**.3487**	0	0	.72
V2	**.3051**	0	0	.63
V3	**.2615**	0	0	.54
V4	0	**.42**	0	.56
V5	0	**.36**	0	.48
V6	0	**.30**	0	.40
V7	0	0	**.4284**	.42
V8	0	0	**.3570**	.35
V9	0	0	**.2856**	.28

Table 2. Factor matrix for hypothetical correlation matrix in table 1.

The number of factors can be as many as the number of variables (though this would simply reproduce the original matrix). The correlation between an indicator and a latent variable is that indicator's *factor loading*. Table 2 depicts three first-order factors (F1, F2, and F3) that account for correlations in the nine manifest variables, and the correlation between the three primary factors (i.e., latent variables) is accounted for by a second-order factor, *g*.

3. Interpretations and Equivocations: *'g'*

Latent variables are sometimes interpreted as conveying some information about cognitive ability or personality. Matters are complicated by the fact that not all latent variables are similarly interpreted; some seem to lend themselves to a realist interpretation more readily than others. SES, for example, is typically not interpreted as real or causally efficacious. A specific value for SES is, depending on one's measurement model, simply a sum-score of a variety of measures including level of education and occupational prestige. Psychological attributes, however, are generally construed to be causally efficacious. To illustrate this point, consider the following two measurement models, which are common in sociological and psychometric research.

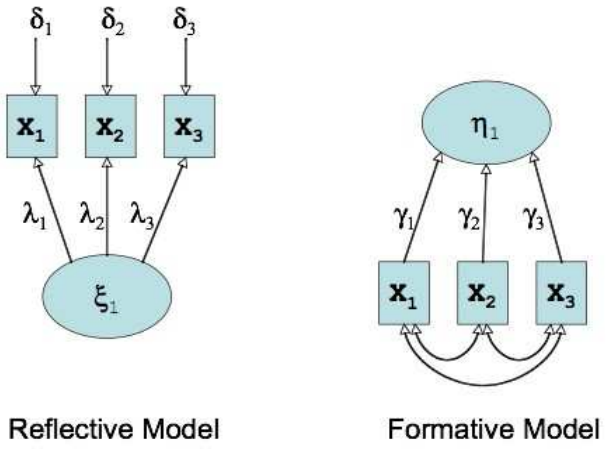


Figure 1. Reflective and formative measurement models.

Appropriating the terminology of Edwards and Bagozzi (2000) and Borsboom (2005), I will refer to the model on the left side of figure 1 as a *reflective* model and to the model on the right as a *formative* model. Each $\mathbf{X_i}$ is a manifest (i.e., observed) variable or indicator such as an item response or test variable. ξ and η are latent variables, each λ_i is the factor loading of each indicator in the left-hand model, and each γ_i is a weight of the indicator with respect to the latent variable.[2] Each δ_i is an error term for the relevant indicator. The reflective model is the typical unidimensional measurement model found in psychometrics. In the measurement of general intelligence, each $\mathbf{X_i}$ would be, for example, a subtest (or item) of a test of cognitive ability and performance on each subtest (or item) would be seen as a function of position on the latent variable *g*; it is differences in positions on *g* which, it is claimed, cause differences in performance on the indicators (hence the direction of the arrows). This is, of

[2] The factor loading of an indicator X with respect to a factor F is an estimated (Pearson) correlation between X and F

course, a simplified model, but it should be sufficient for illustrative purposes. Formative models, on the other hand, are popular in sociological research. For example, socioeconomic status (SES) is often modeled formatively. In the formative model the direction of causal influence is reversed, running from the indicators to the latent variable. The latent variable is regressed on its indicators, not the other way around. One (or a population) occupies a position on SES because of the values of the indicators, such as gross yearly income, and SES is interpreted as summary of the observed measures; no ontological commitment regarding SES independent of its indicators is required. We may even use one's SES score to predict one's level on some unmeasured indicator, but even this does not entail that SES is being treated as existing independently of its indicators.

Some latent variables are generated by variability between persons (interindividual variation) and others are generated by variability within persons (intraindividual variability). *g* represents the former kind of variability. Proponents of *g*-factor models of intelligence cite the robustness of *g* across different factor analytic techniques, biological correlates with *g*, and the apparent impossibility of constructing a test of cognitive ability that does not load on *g* as evidence that there is a single dominant mental ability underlying all cognitive tasks (or at least those sampled by intelligence tests). This is a bit rough since not all intelligence theorists who take *g* to be a requisite explanandum for an acceptable theory of intelligence interpret '*g*' similarly. One source of the heterogeneity in interpretations of '*g*' is confusion over what *g* is. Prominent intelligence researchers sometimes conflate distinct concepts under the name '*g*'. This ambiguity in *g* is not a feature of *g* or factor analysis itself. Rather the ambiguity is a consequence of running distinct statistical concepts together. Offering a

cautionary tale I now turn to a discussion of how various prominent researchers have fallen prey to such confusions.

The four different notions that are sometimes run together under the term '*g*' are

1. *g*-factor: the most general and latent statistical factor that accounts for some portion of the correlation between variables,
2. *g*-score: the weighted sum of an individual's scores on variables that comprise the *g*-factor; i.e., one's position on the latent variable, the *g*-factor,
3. general mental ability: the trait or attribute said to be measured by accepted tests of mental ability which load heavily on the *g*-factor; the purported latent cause of variability in between-subject scores on tests of mental ability,
4. *g*-loading: the correlation between a variable indicating performance (i.e., a variable in a matrix of correlations) and the *g*-factor.

As I will show, running these four related concepts together can lead to serious confusion and odd results.

3.1 Case 1: The Flynn Effect

The Flynn Effect is the well-documented, worldwide steady increase in average IQ (Flynn 1984, 1987, 1999). IQ gains are, on average, 3 points per decade since 1932 (Neisser 1998, 13). Opponents of the centrality of the *g*-factor object that if IQ tests measure mental ability (i.e., *g* in the third sense above) and IQ has been increasing, then so must mental ability. The force of the objection comes from the fact that performance on highly *g*-loaded IQ tests correlates strongly with academic and occupational achievement, but IQ gains have, in fact,

not been accompanied by corresponding gains in academic and occupational achievement (Deary 2001; Flynn 1999), which is a counter-intuitive result given that academic and occupational achievement are correlated with mental ability.

A popular response to this objection to *g*-factor theories of intelligence (Miele 2002; Rushton 1999) is to claim that the IQ gains are hollow in the sense that the gains reflect improved performance on just the non-*g*-loaded sections of the IQ tests. The rationale behind this suggestion is that if the gains in IQ can be accounted for by performance on those sections or items of the test that are not *g*-loaded, then performance is increasing on those sections or items of IQ tests that are not measuring mental ability. This response is an instance of a general strategy for countering the Flynn Effect—to acknowledge that there are IQ-gains, but to deny that there are corresponding gains in general mental ability. This response may seem ad hoc and difficult to reconcile with the fact that IQ-gains are most pronounced on Raven's Progressive Matrices, the psychometric test said to be the "purest" measure of general mental ability, i.e., the Raven's measures general mental ability and little else.

There is another response that follows the aforementioned general strategy, and though it may avoid charges of arbitrariness or ad hoc-ness, the response is marred by equivocation. The response typically goes as follows: if the individual differences, i.e., between-subject variability, in performance on psychometric tests (or the correlations between performance on the tests) has remained constant, then so will *g*. Therefore, IQ gains need not accompany gains in *g*, and since there are no gains in general mental ability, we should expect no gains in achievement. This response equivocates: '*g*' in its first occurrence

only makes sense when interpreted as meaning the *g*-factor (a between-subject statistic), whereas in the second occurrence it is intended in the sense of general mental ability (a within-subject phenomenon).

3.2. Case 2: Psycho-educational Assessment

One dramatic instance of *g*-ambiguity comes from research in psycho-educational research, particularly research on learning disabilities in high-IQ children. The classic operational definition of learning disabilities is a discrepancy between IQ and achievement (Kavale and Forness 1995). In the field of educational psychology there has been a growing concern over misdiagnoses of learning abilities in gifted children. Under the classical operational definition of a learning disability, gifted children are at an increased risk for being diagnosed with learning disabilities—a surprising result. If mental abilities are more differentiated (i.e., less strongly correlated) in populations at the high end of the ability spectrum, then ability and achievement are also more likely to be more discrepant, leading to increased learning disability diagnosis rates. There are reasons to think that gifted children are not at an increased risk for learning disabilities, but that they are at an increased risk for misdiagnoses of learning disabilities (Lovett and Lewandowski 2006).

That mental abilities show greater differentiation at higher levels is well-confirmed (Spearman 1927; Detterman and Daniel 1989; Deary and Pagliari 1991; Detterman 1991; Neisser 1999). For this and other reasons, researchers have questioned the legitimacy of learning disabilities operationalized as IQ/achievement discrepancy. Ability-differentiation in high IQ children entails that IQ tests have lower *g*-loadings for that subpopulation. If *g*-

loading is conflated with general mental ability, the paradoxical conclusion follows that gifted (high ability) children have a lower general mental ability than the rest of the population. Indeed, some psychologists have made this inference on the basis of such this confusion. For example, Ulric Neisser (1999, 131) writes:

> Generally speaking, *g* accounts for less and less of the variance as one moves to individuals with higher and higher scores. This means that more intelligent people have more diverse sets of specific abilities; those in the lower range have not developed those abilities and must use what little they have for virtually every test. Less intelligent people have relatively more *g*!

The first occurrence of '*g*' in the passage quoted above refers to the *g*-factor; however, the second occurrence of '*g*' is intended to refer to general mental ability. Detterman (1991) gestures toward this inference (though he does not make it) when he points out that Spearman may have been committed to the paradoxical result:

> [Spearman] thought that it was *g* that produced the correlations among tests, and that people differed in the *g* they had. Logically, then, groups with the highest correlations among tests should have the largest amount of *g*. Because, in both data reported by Spearman and in my data, the low-IQ groups had the highest correlations among tests, they also must have the largest amount of *g*. In other words, *g* correlates negatively with intelligence, so *g* must be stupidity (1991, 254).

In this passage, the first and second occurrences of '*g*' refer to general mental ability, the fourth occurrence refers to the *g*-loading of the tests administered, and the third and fifth occurrence refers to general mental ability. Conflating these different concepts leads to the paradoxical result that general mental ability is "stupidity." Notice also that the aforementioned paradoxical result entails that IQ tests, purported measures of mental ability, do not measure mental ability, at least not in any meaningful way.

4. Disambiguating *g*

The source of the confusion in the previous two examples is the conflation of levels of mental ability with the variance accounted for by *g*. A person's level of general mental ability, i.e., what is represented by a person's *g*-score, need not be commensurate with the amount of variance that the *g*-factor captures in a collection of scores on mental ability tests. The percentage of variance accounted for (*g*-loading) is not a meaningful statistic for an individual and, hence, cannot indicate an individual's *g*-score or an individual's level of ability. The assumption that the models constructed to represent the structure of variability in the population also fit the individuals within is known as *local homogeneity* (Borsboom 2005). The quoted passages above seem presuppose local homogeneity, and some psychologists explicitly make the assumption as an idealizing principle. Consider the following passage from Anderson:

> Since difference in test scores are the target of explanation, whether these represent differences between 2 adults or longitudinal changes within the same individual

> seems irrelevant. It is taken to be a parsimonious assumption that these differences in scores are to be explained with reference to the same mechanism. Thus, for example, higher synaptic efficiency makes one individual more intelligent, and increasing synaptic efficiency with age makes us more intelligent as we develop (1992, 2).

Nevertheless it makes little sense to appeal to local homogeneity in the context of intelligence research as it conflicts with another commitment typically held by proponents of *g*-factor theories of intelligence, namely the stability of an individual's performance over time on tests of ability. If performance is stable, there is no variability to model; *g* is, by its nature, an interindividual statistic that arises when a diverse battery of mental ability tests are administered to a large population. The claim that the *g*-factor is an indicator of general mental ability in individuals (to a greater or lesser extent) is logically independent from its statistical nature. If *g* has any empirical significance, it is as an indicator that individual differences in test scores can be modeled along a single dimension. Bartholomew (2004) likens *g* to longitude in this respect. Longitude is a dimension along which geographical points can differ. The Eiffel Tower and Lenin's sarcophagus, for example, differ along this dimension. Each of these objects occupies a point on this dimension and their locations can be partially described by specifying their "score" on that dimension. However, the location of the Eiffel Tower on its own is insufficient for constructing the dimension of longitude as the Eiffel Tower stays put.

5. Conclusion

I have drawn attention to several problems in the conceptual foundations of psychological measurement. First I argued for a pragmatic distinction between latent and manifest variables. This account avoids the controversy concerning the so-called "observable/theoretical" distinction associated with the realism debates. I then turned to equivocal notions of the *g*-factor. The diagnosis of the ambiguity is that it results from a tacit assumption of local homogeneity, which, in turn, leads to a conflation of several distinct statistical and psychological concepts. Once teased apart, no paradox arises; however, echoing others including Borsboom (2005), I noted that if *g* denotes general mental ability, it seems that no individual has it, for *g* is an interindividual statistic. As a dimension of variability, psychometricians can assign individuals scores on that dimension, though doing so involves substantive empirical assumptions and merely provides an ordinal standing of that individual relative to the relevant population.

References

Anderson, M. 1992. *Intelligence and Development*. Cambridge: Blackwell.

Bartholomew, D. 2004. *Measuring Intelligence: Facts and Fallacies*. New York: Cambridge.

Borsboom, D. 2005. *Measuring the Mind: Contemporary Issues in Psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D. 2008. "Latent Variable Theory." *Measurement: Interdisciplinary Research and Perspectives* 6: 25-53.

Deary, I. J., and Claudia Pagliari. 1991. "The Strength of g at Different Levels of Ability: Have Detterman and Daniel Rediscovered Spearman's 'Law of Diminishing Returns'?" *Intelligence* 15: 247-250.

Detterman, D. K. 1991. "Reply to Deary and Pagliari: Is g Intelligence or Stupidity?" *Intelligence* 15: 251-255.

Detterman, D. K., and Mark H. Daniel. 1989. "Correlations of Mental Tests with Each Other and with Cognitive Variables are Highest for Low IQ Groups." *Intelligence* 13:349-359.

Edwards, J. R. and R. P. Bagozzi. 2000. "On the Nature and Direction of Relationships between Constructs and Measures." *Psychological Methods* 5:155-174.

Flynn, J. R. 1984. "The Mean IQ of Americans: Massive Gains." *Psychological Bulletin* 95:29-51.

Flynn, J. R. 1987. "Massive IQ Gains in 14 Nations: What IQ Tests Really Measure." *Psychological Bulletin* 101:171-191.

Flynn, J. R. 1999. "Searching for Justice: The Discovery of IQ Gains Over Time." *American Psychologist* 54:5-20.

Jensen, A. 1998. *The g Factor*. Westport: Praeger.

Kavale, K. A., and S. R. Forness. 1995. *The Nature of Learning Disabilities: Critical Elements of Diagnosis and Classification*. Mahwah, NJ: Erlbaum.

Lovett, B. J., and L. J. Lewandowski. 2006. "Gifted Students with Learning Disabilities: Who are They?" *Journal of Learning Disabilities* 39:515-527.

Miele, Frank. 2002. *Intelligence, Race, and Genetics: Conversations with Arthur Jensen*. Boulder, CO: Westview Press.

Neisser, U. ed. 1998. *The Rising Curve*. Washington DC: American Psychological Association.

Neisser, U. 1999. "Two Views About the g Factor: The Great g Mystery." *Contemporary Psychology APA Review of Books* 44:131-133.

Rushton, J. P. 1999. "Secular Gains in IQ Not Related to the g Factor and Inbreeding Depression Unlike Black-White Differences: A Reply to Flynn." *Personality and Individual Difference* 26:381-389.

Sesardic, Neven. 2000. "Philosophy of Science That Ignores Science: Race, IQ and Heritability." *Philosophy of Science* 67:580-602.

Sobel, S. E. 1994. "Causal Inference in Latent Variable Models." In *Latent Variables Analysis*, ed. A. von Eye and C.C. Clogg, 3-35. Thousand Oaks: Sage.

Spearman, Charles. 1927. *The Abilities of Man: Their Nature and Measurement*. New York: Macmillan.

Trout, J. D. 1998. *Measuring the Intentional World: Realism, Naturalism, and Quantitative Methods in the Behavioral Sciences*. Oxford: Oxford University Press.

Effects and Artifacts: Robustness Analysis and the Production Process

Abstract: Scientists often use multiple independent methods of identification to distinguish reliable results from those produced in error (artifacts). This process is referred to as 'robustness analysis'. I argue that even though robustness analysis is useful for differentiating natural phenomena from artifacts, it fails to differentiate experimentally produced effects from artifacts. I argue that to bypass this problem, we can re-frame the role of robustness analysis to focus on cross-comparison between methods of production. Focusing on the production relation provides information about *how* changes in conditions alter given effects, without first having to make a distinction between effect and artifact.

1. Introduction.

Scientists often use multiple methods of identification to distinguish reliable results from those produced in error. When methods converge on an object or process, inferences are made about reliability. For example, Perrin (1913) successfully used multiple independent methods of measurement and inference to converge on Avagadro's number, thus supporting its objectivity. Unsuccessful convergence (or divergence) indicates error. Recently, disagreement has surfaced about the failure of reproducing results about an arsenic-consuming living organism (Reaves et al. 2012). In 2010 a novel discovery seemed to redefine how biologists understand the chemistry of living organisms by questioning whether phosphorous is necessary for cellular function. A bacterium in the arsenic-rich waters of Mono Lake was found. Under a set of specific experimental conditions the organism was found to replace phosphorous with arsenic in its DNA (Wolfe-Simon et al 2011). However, using more stringent, *independent* experimental conditions to eliminate phosphorous and to "purify" the DNA samples of any clinging arsenate, Reaves et al. (2012) did not find covalently bound arsenate in the DNA structure. The divergence in results indicates that there was an experimental error produced and that the original arsenic-consumption effect was an artifact of the lack of purification in the preparatory procedure.

This method of identification based on converging results is commonly referred to as 'robustness analysis'. It is a methodological process of generating conclusions or results that converge over a variety of independent identifications, models, measurements, or derivations

(Wimsatt 2007, 43). Philosophers of science have discussed robustness analysis in modeling[1] as well as in experimentation and evidence.[2] I focus on the latter. According to Wimsatt, robustness analysis grounds realism, reliability, and objectivity and distinguishes the ontologically and epistemologically "trustworthy" from what is unreliable (2007, 56). Specifically, it differentiates real objects, events, and processes from "artifacts", which are results produced in error (2007, 38).[3] The characteristic of "artifactual" results is that they are unstable in the context of multiple independent methods of identification (Wimsatt 2007, 56). To understand how stability/convergence and instability/divergence works, a bit of detail about the process of robustness analysis is necessary.

In robustness analysis, several independent methods converge on a common consequence despite independent conditions, abstractions, and idealizing assumptions (Levins 1966; Wimsatt 2007). This consequence, often referred to as a 'robust consequence', can be a prediction, property, or result.[4] After the robust consequence is generated, a 'robust

1 See Levins (1966); Glymour (1980); Weisberg (2006); and Wimsatt (2007).

2 See Horwich (1982); Hacking (1983); Franklin (1997); Sober (1989); Cartwright (1991); Trout (1998); Culp (1994); Woodward (2006); Stegenga (2009).

3 In this discussion I use 'artifact' in reference to experimental artifacts, which are results produced in error. Technological artifacts will not be addressed except for in the development of a characterization of 'experimental artifacts' in Section 2.

4 Levins and Wimsatt explicitly characterize common consequences and robust theorems in relation to the modeling process. Here, I extend 'consequences' to any method of

theorem' is created, which states the robust consequence *relatively* independent of the different conditions, abstractions, and idealizing assumptions.[5] Wimsatt (2007) frames robustness analysis as an effective "heuristic" that can be used to show how a robust consequence does not depend on the different details of each method (56). These details are unstable between methods and thus their divergence fades into the background, while the robust consequence is set into focus. Wimsatt (2007) uses the example of detecting properties of planets with multiple independent imaging techniques. According to Wimsatt, if a given signal is weak and the noise of each imaging technique is strong, by combining techniques, the signal strength will increase while the different types of noise will not (2007, 57). The noise is independent, random, and differs between methods, while the signal is invariant and over each method.

At first glance, philosophical examples of robustness analysis, signal detection (Wimsatt 2007; Campbell 1966), ecological populations and species polymorphism (Levins 1966), predicting predator-prey relations using Lotka–Volterra equations (Weisberg 2006), climate change (Lloyd and Parker 2009; Lloyd 2010) do not seem to have anything particular in common about the *type of regularities* studied. They show diverse application of robustness analysis to physical and biological regularities. But these are all examples of using robustness analysis to distinguish *natural phenomena* from results produced in error. That is,

---

indentification.

5 In discussions about the robust theorem, Weisberg (2006) places attention on underlying common structure. Levins (1966) and Wimsatt (2007) discuss the falsifying idealizations.

in these types of examples objects, events, and processes are *discovered* by carefully comparing multiple independent methods of identification. To use a signal detection analogy, methods are fine-tuned so that the signal of a given regularity is made clear. However, there is a different type of regularity that requires careful differentiation and attention: effects that are experimentally *produced* rather than discovered.

Hacking distinguishes 'phenomena' and 'effects'. He characterizes 'phenomena' as "observable regularities" (1983, 221). These are regularities that are not the result of experimental intervention—e.g., the planets and stars. According to Hacking, there are few phenomena in nature waiting to be observed but science is full of regularities that are produced through intervention as 'effects' (1983, 227).[6] He distinguishes the two types of regularities:

> Phenomena and effects are in the same line of business: noteworthy discernible regularities. The words 'phenomena' and 'effect' can often serve as synonyms, yet they point in different directions. Phenomena remind us, in that semiconscious repository of language, of events that can be recorded by the gifted observer who does not intervene in the world by who watches the stars. Effects remind us of the great experiments after whom, in general, we name the effects: the men and women, the Compton and Curie, who intervened in the course of nature, to create a regularity which, at least at first, can be seen as regular (or anomalous) only against the further background of theory. (Hacking 1983, 224-225)

6 Often, Hacking uses 'phenomena' for 'effects'.

Effects require carefully planned production conditions. According to Hacking, the aim of experiments is to "create," "refine," and repeat the effects produced in an experiment (1983, 229-230). But because effects fall apart when conditions are modified, it is likely that effects are only produced under *specific* conditions in an experimental setting (1983, 225-226). Suppose that we apply robustness analysis to a given effect to figure out if it is reliable. We develop independent methods for the production of this effect, but because effects are sensitive to conditions, convergence is unsuccessful. What can we conclude about reliability? Using diverging results will not help. It has been argued that divergence or "discordance" in results thwarts useful inferences about reliability and error (Stegenga 2009). An important methodological and epistemological question arises for robustness analysis: *If effects require careful experimental production and are sensitive to changes, how do we know when an effect is genuine, rather when it is a result produced in error (an experimental artifact)?*

In this discussion I argue that robustness analysis does not reliably differentiate between effects and experimental artifacts, but that we can modify its function in such instances to give us useful information about the production relation. In Section 2, I argue that robustness analysis fails to differentiate genuine effects from artifacts because: 1) "Arrangements" cannot be differentiated into the "real" and the artificial; and 2) Introducing multiple methods will change the conditions, thus producing diverging results. In Section 3, I argue that to bypass this problem about differentiating effects from artifacts, we can re-frame the role of robustness analysis. I propose that we can use cross-comparison to understand

how changes in conditions alter given effects. This allows us to understand the experimental production *process*.

2. Differentiating Effects from Experimental Artifacts.

In this section I argue that both effects and experimental artifacts are condition-sensitive, and for this reason difficult to distinguish. Hacking illustrates the condition sensitivity of effects by describing the original Hall effect experiment, where an electric current is passed through a gold leaf in the presence of a perpendicular magnetic field. These conditions produce a potential difference across the conductor (the leaf) and at right angles to the magnetic field and conductor (1983, 224). Hacking says that even though the conditions were carefully planned and the apparatus was human-made, we have the intuition that the phenomenon was "discovered" in the laboratory rather than created (1983, 225). But according to Hacking, the "arrangement" of conditions behind the Hall effect only occurs in the laboratory. He says, "I suggest, in contrast, that the Hall effect does not exist outside of certain kinds of apparatus. Its modern equivalent has become technology, reliable and routinely produced. The effect, at least in a pure state, can only be embodied by such devices" (1983, 225). Hacking's analogy between technological effects and experimental effects provides an important point. Both types of effects are sensitive to the manipulations of the arrangement of conditions. Kroes (2003) provides a useful characterization that

supplements Hacking's analogy between experiment and technology, and he uses it specifies 'experimental artifacts'.[7]

Kroes (2003) characterizes 'artifacts' in general as resulting from intentional human action and directed toward a specific function. However, he points out that artifacts "obey the so-called laws of nature; that is, their behavior can be explained causally in a nonteleological way" (2003 19). We can summarize Kroes' (2003) points into two important features of artifacts: 1) Human design/specified function; and 2) Regularities explainable by laws of nature. Kroes' (2003) initial characterization focuses on technological artifacts, which require a considerable element of human structural design and function. For example, the structured interactions of thin film transistors (TFT's) can be manipulated for the purposes of LED technology (Machrone 2013).[8] I add that such technological artifacts often result from experimentally produced effects.[9] If technological artifacts are products of human design then

7 Kroes's (2003) is noteworthy because it is an extension of Hacking's argument for experimentation. Chakrabarty (2012) provides an important summary of definitions of artifacts.

8 This technology is based on the Hosono et al. (2005) research on crystal structure.

9 Hosono et al. (2005) found that by manipulating the crystal structure of certain materials we can experimentally produce a compound that conducts electricity. The reason why this is an experimental effect is because even if a given compound's conductivity is low (e.g., due to the asymmetry in the crystal structure), adding titanium atoms to its structure produces symmetric cages, which allows free electron flow (Hosono et al. 2005).

experimental artifacts can be characterized as design products of the apparatus and the arrangement of measurement conditions. Kroes (2003) makes a fine-grained characterization of experimental artifacts by discussing "artificial environments" vs. the "object system" of study. Drawing on Franklin's (1986) discussion he characterizes experimental artifacts as "results that are generated by the artificial environment or artificial means of observation of the natural phenomena under study" (2003, 71). In his discussion of distinguishing artifacts from genuine effects he suggests:

> The results of an experiment are always the outcome of the object system interacting with an artificial environment, and therefore it is always necessary to filter out the component in the results that tells us something about the object system. (2003, 71)

While this suggestion is useful in distinguishing phenomena from artifacts, it is not helpful for distinguishing effects from artifacts. The reason why is because it assumes that we can filter out the error in an experimental arrangement by separating the artificial environment from the object system. Sometimes, experimental arrangements do not lend themselves to filtering because they heavily rely on the manipulation of total conditions rather than on the distinction between artificial and natural conditions. Take Hasok Chang's (2004) discussion the manufacturing process of fixed points in thermometry.

Chang details how the boiling point of water varies with differences in atmospheric pressure and dissolved gas (Chang 2004, 15-19). Different arrangements of conditions will

produce a different boiling point. The effect of boiling point is so sensitive to the manipulation of conditions that water can boil at 101.9 degrees C merely in the presence of dissolved gas (Chang 2004, 19). In the history of fixed points like the boiling point of water, material conditions have to be fine-tuned to "manufacture" fixity (2004, 49). Here, Kroes's (2003) suggestion to filter out the "artificial environment" and focus on the "object system" is not helpful. In the case of boiling point, all we have is "arrangements" that are manipulated. Sometimes nature manipulates the same effects that scientists create in the lab. For example, Hacking says, "If anywhere in nature there is such an arrangement, with no intervening causes, then the Hall effect occurs" (1983, 226). So we can't claim that some arrangement themselves are artificial and some are natural. Additionally, in the lab, differentiating which arrangements are artificial and which ones are natural is just as difficult. Suppose that we manipulated atmospheric pressure but let dissolved gas run out naturally. While the former requires human intervention it is unclear if it is artificial. The latter requires no human intervention but is still changing given the conditions in the room. In both instances there is a change in conditions relevant to the production of the effect, but the division between artificial and object system is unclear. Perhaps multiplying production methods will help.

A common specification in robustness analysis is to use multiple *independent* methods.[10] Suppose that we want to check to see if water boiling at 101.9 degrees C is an

10 Philosophers such as, Nagel (1939); Horwich (1982); Franklin (1984); Sober (1989); Trout (1993); Culp (1994); Keeley (2002); Staley (2004); Douglas (2004); Novack (2007);

artifact. We repeat the exact same conditions and reproduce the effect of water boiling at 101.9 degrees C. Here, we would not be using independent methods. For example, in our reproduction of the boiling point, the physical conditions are the same types of conditions. In fact, because of the sensitivity of boiling point they have to be the same types of conditions to reproduce the same boiling point. If we change the physical conditions, results will diverge from those of our initial condition arrangement. The point is that using independent methods of measurement with sensitive effects will produce diverging results. The divergence does not differentiate effects from artifacts.

The reason why robustness analysis fails to differentiate effects from artifacts is important. In the context of "discovered phenomena," condition sensitivity is precisely the indicator in robustness analysis that tells us when something is produced in error. The mesosome is an example of a cellular structure that appeared in multiple types of microscopy measurement methods. But it was later found to be a result produced in error by chemical fixation in a specific preparatory procedure.[11] In the context of the mesosome, as soon as we switch preparatory procedures the mesosome disappears (Rasumussen 1993; 2001). This indicates that it was a result produced by the preparatory procedure. In the context of so-

---

Wimsatt (2007); and Stegenga (2012), have discussed conditions defining independence. I will not summarize the differences in views here.

11 For debate on measurement methods and the mesosome see Culp (1994), Rasmussen (1993; 2001), and Hudson (1999). Stegenga uses this example to illustrate converging results can support false conclusions (2009, 653).

called arsenic-ingesting bacteria, as soon as we carefully "wash" the DNA structure we see that there is no covalently bound arsenate in the DNA sample, and so the bacteria does not replace phosphorous with arsenate in its DNA structure (Reaves et al. 2012). This indicates that the pre-spectroscopy DNA "purification" procedure produced the result. In these contexts there is a phenomenon that is *independent of the arrangement of conditions in the preparatory procedures*. In the case of effects, the thing produced is arrangement-dependent. Given that artifacts are arrangement-dependent also, we need another condition that differentiates that two. But instead I propose that we look past the distinction to learn something unique about robustness analysis in the context of effects.[12] We can use diverging results to understand *how* changes in conditions alter given effects.

3. Production Analysis.

How can diverging results provide useful information in the context of effects? Robustness analysis focuses on converge. When using the analysis, we focus on consequences common to several independent methods. But there are two philosophical accounts of robustness analysis that say something useful about diverging results (Weisberg 2006; Keyser 2016). I take elements from each in order to develop 'production analysis'.

Weisberg (2006) focuses on predictions rather than consequences and models rather than methods, but we can modify his steps to be useful for experimental production.

12 In the remainder of the discussion I use 'effects' only in reference to things produced by the arrangement of experimental conditions.

Weisberg outlines robustness analysis using four steps. First, we find a robust property, which consists of finding a property, experimental result, or prediction that is common to a set of models with different idealizing assumptions (2006, 736). In the context of effects, we do not *discover* robust properties. Rather, we *produce* certain regularities. As was presented in Section 2, because of the sensitivity of effects and the requirement of independence, effects will fail to converge between production methods. Weisberg's (2006) second step is to investigate the "common structure" by looking at the common features of the models that give rise to the robust property (2006, 737)." In the context of effects, we can look at this as the common features of production that create a given effect. However, if we have diverging results this step is not useful. The third step is an "empirical interpretation" of the mathematical structures from step two. That is, according to Weisberg in the third step of robustness analysis we are concerned with "interpreting the mathematical structures as descriptions of empirical phenomena (2006, 738)." This step is uninformative for produced effects because we are not using modeling methods to link a theory to a natural phenomenon, but rather we are using experimental methods to manufacture the effect itself. In the context of effects, our concern is a *production relation* rather than a representation relation. However, the final step of Weisberg's (2006) analysis is important. Weisberg says, "Finally, the theorist can conduct stability analysis of the robust theorem to determine what conditions will defeat the connection between common structure and robust property" (2006, 738).

We can transform Weisberg's stability analysis into what I call a 'production analysis'. Instead of looking at what model conditions will defeat the robust property, we can look at *how* specific experimental conditions produce differences and similarities in effects. This requires structuring production analysis into features. First, multiple methods of production are involved. These can be referred to as 'production processes'. The production processes will contain experimental conditions (i.e. "arrangements") that are causally relevant to the effects they produce.[13] These methods do not have to be independent. For example, we can compare multiple production processes of the boiling point of water, using the same condition parameter values. In fact, when it comes to effects, it may help to have both dependent and independent methods. The reason why is because we can track what kinds of changes occur from similar processes as well as what kinds of changes occur from different processes. Second, conditions in production processes are compared to map out convergence and divergence. It may be that all production processes diverge (or converge), or that there is a mix of convergence and divergence. This feature of production analysis contains not only a comparison of convergence and divergence but also a comparison of the *conditions* in each production process. Third, theory is applied to the two levels of comparison to explain why certain conditions produce (or fail to produce) specific changes in the effect. This final component of production analysis is informative about *how* conditions change effects. To detail how production analysis works, we can

13 A particular theory of causation is not important for our purposes.

draw on elements from Keyser's (2016) discussion about cross-comparison and theory in robustness analysis.[14]

For his account of robust measurement, Keyser (2016) assumes an important mechanism discussed by Woodward (2003). Manipulating one variable to see changes in another is causally informative. Keyser (2016) argues that theory in the presence of diverging measurement results can explain *why* divergence occurs. He uses the example of mixed convergence and divergence in multiple modes of temperature measurement. When multiple thermometers converge but others diverge, theory steps in to analyze the conditions behind the divergence (Keyser 2016, 10). Keyser proposes that theory homes in on specific physical differences in thermometers—e.g., the liquid used in a given thermometer—in order to explain how those features produce differences in results (2016, 10-11). This differential comparison and explanation process is useful for measurement in terms of specifying the "location" of error (Keyser 2016, 11). But in the context of production, what does it mean to have an explanation about why divergence occurs?

To be more informative, I add that diverging results can provide information about *which* conditions produce a change (or fail to produce a change) in the effect. In other words, divergence *locates specific changes* in the production relation between condition and effect. Then theory can be used to explain *why* those production changes occur. Suppose that there are two experimental setups (or production processes) for the boiling point of water. In the first setup, there is no presence of dissolved gas. In the second setup, we add a certain

14 Cited with permission from author.

amount of dissolved gas, which increases the temperature by a certain amount of degrees. Comparing the two setups in terms of divergence and also in terms of their conditions, we see that there is divergence in results and that the difference in conditions is in the presence of dissolved gas. Each production process creates two different boiling points. Without the presence of theory these comparisons are uninformative. But with the presence of theory we understand that the difference between the two production processes can be explained by a specific theoretical reason (e.g., we are influencing “vapor pressure” in different ways in each process).

To summarize, production analysis requires: 1) Two levels of comparison, which locate what conditions are relevant to the effects. Effects are compared to see the presence of divergence and convergence. Conditions are also compared to see what differences may be responsible for changes (or failures of change) in effects; 2) Theoretical explanation about why divergence is being produced and what conditions are responsible. In production analysis the focus is on how certain conditions produce effects. The aim is to understand *production relations* in the context of multiple production processes. The benefit of production analysis is it provides useful information about the production process without first having to differentiate effect from artifact.

4. Concluding Remarks.

While robustness analysis is informative for the distinction between phenomena and results produced in error, it fails to distinguish experimentally produced effects from artifacts.

The reason why is because in the context of production, both effects and artifacts are sensitive to changes in methods of production. This means that diverging results will not be informative for differentiating effects from artifacts. Condition sensitivity is precisely the indicator in robustness analysis that tells us when something is produced in error; and both effects and artifacts sound off this indicator. I argued that to bypass this problem about differentiating effects from artifacts, we can re-frame the role of robustness analysis to focus on: 1) Cross-comparison between results and also between conditions; and 2) Theoretical explanation of that cross-comparison. I refer to this process as 'production analysis'. I proposed that we use diverging results to understand *how* changes in conditions alter given effects. This provides information about the production relation and a new role to robustness analysis in accounting for how conditions are relevant to effects.

References

Cartwright, N. 1991. "Replicability, Reproducibility, and Robustness – comments on Harry Collins." *History of Political Economy* 23: 143–55.

Chakrabarty, Manjari. 2012. "Popper's Contribution to the Philosophical Study of Artifacts." In: *[2012] Philosophy of Science Assoc. 23rd Biennial Mtg (San Diego, CA) > PSA 2012 Contributed Papers*.

Chang, Hasok. 2004. *Inventing Temperature.* Oxford University Press.

Culp, S., 1994. "Defending Robustness: The Bacterial Mesosome as a Test Case." in

D. Hull, M. Forbes and R. Burian (eds.), *PSA 1994*, East Lansing: Philosophy of Science Association, 47–57.

Douglas, Heather. 2004. "The Irreducible Complexity of Objectivity." *Synthese* 138(3): 453-473.

Franklin, Allan. 1986. *The Neglect of Experiment*. Cambridge, Massachusetts: MIT Press.

———1997. "Calibration." *Perspectives on Science* 5:31-80.

Glymour, Clark. 1980. *Theory and Evidence*. Princeton, NJ: Princeton University Press.

Hacking, I. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.

Horwich, Paul. 1982. *Probability and Evidence.* Cambridge: Cambridge University Press.

Hosono, H., Hirano, M., Ota, H., Kamiya, T., Nomura, K. 2005. "Amorphous Oxide And Thin Film Transistor." *US Patent App* 10:592.

Hudson, R. G., 1999. "Mesosomes: A Study in the Nature of Experimental Reasoning." *Philosophy of Science* 66:289–309.

Keeley, Brian. 2002. "Making Sense of the Senses: Individuating Modalities in Humans and Other Animals." *The Journal of Philosophy* 99(1): 5-28.

Keyser, Vadim. 2016. "A New Theory of Robust Measurement." *American Philosophical Association 2016 Pacific Division Meeting*. Pre-print:

http://www.apaonline.org/members/group_content_view.asp?group=110424&id=476093

Kroes, Peter. “Physics, Experiment, and the Concept of Nature.” In Radder, Hans, ed.. 2003. *The Philosophy of Scientific Experimentation*. Edited by Hans Radder. University of Pittsburgh Press.

Levins, R. 1966. “The Strategy of Model Building in Population Biology.” *American Scientist* 54: 421–431.

Lloyd, E. A. 2010. “Confirmation and Robustness of Climate Models.” *Philosophy of Science* 77, 971–984.

Lloyd, E. A., Parker, W. 2009. “Varieties of Support and Confirmation of Climate Models.” *Proceedings of the Aristotelian Society* lxxxiii: 1467-8349

Machrone, Bill. 2013. “Powering the Resolution Revolution.” *The Wall street Journal*. http://online.wsj.com/ad/article/vision-igzo

Perrin, J. 1913. *Les Atomes*. Paris : F. Alcan. (Atoms. by D. Li. Hammick, Trans., 1916). New York: D. Van Nostrand. (Reprinted Kessinger Publishing, 2007).

Rasmussen, N. 1993. “Facts, Artifacts, and Mesosomes: Practicing Epistemology with the Electron Microscope.” *Studies in History and Philosophy of Science* 24: 221–265.

———2001, “Evolving Scientific Epistemologies and the Artifacts of Empirical Philosophy of Science: A Reply Concerning Mesosomes.” *Biology and Philosophy* 16: 629–654.

Reaves, M.l., Sinha, S., Rabinowitz, J.D., Kruglyak, L. Redfield, R.J. 2012. "Absence of detectable arsenate in DNA from arsenate-grown GFAJ-1 cells." *Science* 337(6093):470-3.

Sober, Elliot. 1989. "Independent Evidence About a Common Cause." *Philosophy of Science* 56: 275-287.

Staley, Kent. 2004. "Robust Evidence and Secure Evidence Claims." *Philosophy of Science* 71:467-488.

Stegenga, Jacob. 2009. "Robustness, Discordance, and Relevance." *Philosophy of Science* 76: 650-661

——— 2012. "Rerum Concordia Discors: Robustness and Discordant Multimodal Evidence." in *Characterizing the Robustness of Science* Boston Studies in the Philosophy of Science, Springer, 207-226.

Trout, J.D. 1998. *Measuring the Intentional World*. Oxford: Oxford University Press.

Weisberg, Michael. 2006. "Robustness Analysis." *Philosophy of Science* 73, 730–742.

Woodward, J. 2003. *Making things happen: A theory of causal explanation*: Oxford University Press, USA.

——— 2006. "Some Varieties of Robustness." *Journal of Economic Methodology* 13:2, 219-240.

Wimsatt, William. 2007. *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge: Harvard University Press.

Indexically Structured Ecological Communities

Abstract. Ecological communities are seldom, if ever, biological individuals. They lack causal boundaries as the populations that constitute communities are not congruent and rarely have persistent functional roles regulating the communities' higher properties. Instead we should represent ecological communities *indexically*, by identifying ecological communities via the network of weak causal interactions between populations that unfurl from a starting set of populations. This precisification of ecological communities helps identify how community properties remain invariant, and why they have robust characteristics. This is a more general framework than individuality, respecting the diversity and aggregational nature of these complex systems while still vindicating them as units worthy of investigation.

Word Count: 4955

1. Introduction

Ecology studies the distribution and abundance of populations across landscapes and over time. Community ecology has long operated with the assumption of "local determinism": i.e. that ecological patterns are primarily explained by the interaction of local populations within a community. The 'local ecological community' functions as the core unit of investigation; it is thought to have discrete boundaries, stable composition, predictable dynamics over time, and allows for inferences made from one community to apply to the next. But there have been many dissenting voices within the ecological research tradition who instead argue for *ecological individualism*, emphasizing that populations generally move around a landscape of their own accord driven by chance and by abiotic factors and are not heavily influenced by their local neighbours and as a result ecological communities are largely ephemeral compositions of populations. This debate drives considerations whether there are law-like regularities in community ecology.

To arbitrate this debate philosophers and biologists have provided an analyses of the conditions for an assemblage- a collection of populations in a space- to be an ecological community. Namely, assemblages should be a biological individual just like an individual organism or a lineage. If an ecological community is a biological individual then it is the cohesive and distinct entity that local determinism presupposes. Jay Odenbaugh and Kim Sterelny independently specified the conditions under which an ecological assemblage can be thought of as an objective and important unit in nature, an *ecological community.* Both authors leave it empirically open as to whether and which assemblages satisfy the conditions they present (Odenbaugh, 2007; Sterelny; 2006).

I argue that as ecological communities so rarely satisfy these conditions we need an alternative account of ecological communities. Instead ecological systems are largely

aggregations of individual populations unlinked by stable, strong causal interactions. As a result they are better described indexically, as causal networks which unfurl from a specific point of reference. This acts to fix the reference in these unsystematic systems and allows for the identification of the robust parts and robust properties of ecological systems. To infer from one community to the next we need a precise account of the identity of the units we are discussing. This elaboration on indexical communities provides that. This is not to say that ecological communities will never be biological individual, there will be limiting cases. But these lie so far from the norm that we need a framework that better represent the degree of variation in ecological assemblages.

This proposal provides a substantively different framework to biological individuality, diversifying the ontological toolkit of philosophers of biology. To do this I will first introduce the theory behind ecological communities as biological individuals. I then elaborate on what it takes for a collection of parts to be an individual and why ecological systems are almost always not. Into this lacuna I then present the indexical account of ecological communities and the advantages it entails.

2. Communities as Biological Individuals

Multicellular biological individuals evolved from single cell biological individuals. While multicellular individuals often evolve from a single species population it is not uncommon for multispecies assemblages to form individual organisms such as lichens[1]. But there is also integration without unification; most large metazoans only function by inheriting bacteria that maintain them in a symbiotic relationship (perhaps also some biofilms). These biological

[1] Single species populations transitioning into an individual are referred to as fraternal transitions. Multispecies transitions are referred to as egalitarian transitions (Queller, 2000).

phenomena demonstrate that strong ecological interactions precede transitions in individuality and multispecies assemblages can be individuals. But where does this leave the assemblage's community ecologists are familiar with? They are clearly less of an individual than populations in close symbiotic relationships. Any account of biological individuality for communities needs to be able to account for degrees of individuality. One way to indicate this is by providing a set of conditions which if fulfilled rightly counts an ecological community as an individual, then leave it open as to whether any actual ecological community satisfies these conditions. This is what both Sterelny and Odenbaugh do. The conditions they present follow:

2.1. Boundaries

Individuals, as spatio-temporal entities comprised of interacting sub-parts, have boundaries. For interacting parts to be a whole there must be strong causal interactions creating internal cohesion within the system which isolates it from external influences. The system parts in community ecology are the populations which causally interact, creating feedback loops maintaining local populations and excluding external populations from invading the local system. Sterelny particularly notes that local niche construction is one way populations can maintain an assemblage. Famously, Australian plants including Gums, Banksias, and Melaleucas are adapted to fire and facilitate the presence of each other by making their local environment more fire prone. Under this conception of boundaries ecological communities are bound by interaction strength between populations (Levins and Lewontin, 1985). While this does not necessarily mean that populations in the system will be congruent, strong causal interaction is associated with spatial overlap so congruence of community populations is expected.

2.2. Internally Structured

The populations that belong to an ecological individual should act in ways that police the identity of that individual. Interspecific interactions- such as predation, competition, and, mutualism- are thought to form a lattice of positive and negative feedback loops, regulating the community and creating stability. When you couple these interactions with stable geographic ranges of the populations you gain a picture of stable economy of nature in which there is persistence of local population identity due to the specific roles that these populations play. Internal structure is the product of both feedback loops that act to maintain population identity in an area and the persistence of specific populations playing particular roles in this local community.

2.3. System-level Properties

If we wish to include local ecological communities in our general scientific ontology there has to be a reason to talk about communities rather than just talking about the populations that make up communities. There should be predicates and properties which are needed for describing phenomena at the community level. System-level properties are an explanadum to be explained by the assemblage and an explanans for ecological and evolutionary hypotheses. Properties generally discussed on the community level are associated with the maintenance of multi-species interaction networks (*community network structures*), the maintenance of composition identity or aggregative features (*emergent community properties*), or the various material outputs that the joint assemblage create (*community outputs*). Odenbaugh treats system-level properties as necessary for community existence: 'species populations form an ecological community just in case… they possess a community level property' (pg 636). He primarily mentions interspecific interactions and the feedback loops they create as community level properties. Sterelny describes emergent community properties, identifying several candidate emergent properties from the diversity-stability hypothesis such as

community population stability and community biomass production. The productivity and abiotic features ecological communities produce have become an area of keen interest for conservation science. Many ecologists have attempted to justify the preservation of ecological communities by appealing to the 'ecosystem services' - capacities commonly attributed to the community as a whole - which they provide. These system-level properties feature in ecological explanation and therefore need to be able to be represented by an account of ecological communities.

3. Problems with Individuality

Sterelny- Odenbaugh individuality features a tripartite criteria that ecological communities need to fulfill: they should be bound causally, they should have internal regulation, and they should have system-level properties. Sterelny represents these criteria hyper-dimensionally noting that each can be more or less instantiated (see fig 1). This is partially true, but these axes are not independent, as both authors independently note (Maclaurin and Sterelny, 2008; Odenbaugh, Forthcoming). Internal regulation demands boundaries contain regulatory patterns of interactions and system-level properties require the populations to be structured. This view implies that if communities do not have boundaries they will not have internal regulation and without internal regulation they will lack system-level properties. While I argue that communities do not have robust boundaries and their internal structure is not as stable as individuality requires, I maintain that ecological systems can have system-properties. Loose aggregative ecological systems produce system-level properties by what I call machine robustness and ensemble robustness.

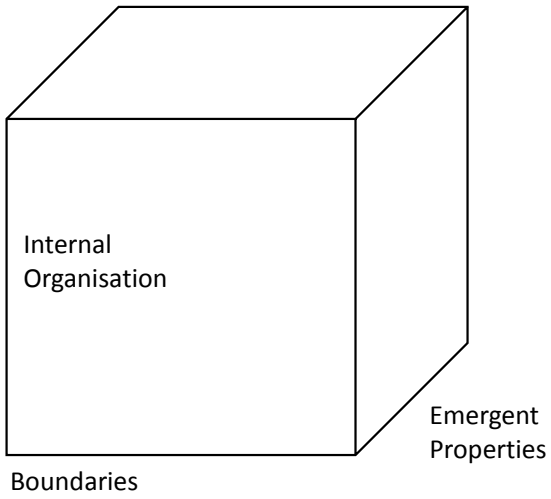


Figure 1. *Sterelny's Multidimensional Representation of an Ecological Individual.*

The best way to identify both the relation between robust outputs and the stability of the system that produces these outcomes can be found in Bill Wimsatts conceptual framework of *multiple decomposition*. For community boundaries to be 'real' they should be *descriptively robust.* An entity is descriptively robust when multiple interventions, multiple types of intervention, and different descriptions of relevant properties pick out organisation that is largely congruent (Wimsatt; 2007). The parts described by a 'decomposition' from one theoretical perspective largely align with those describe from another. By applying this procedure to local ecological communities, as I will now show, we find that communities as we often have understood them are not descriptively robust as a whole and therefore are not biological individuals.

We multiply decompose a 'local ecological community' by identifying the causal system that the different co-located populations belong to. Each population will belong to an

ecological system consisting of just those populations to which they are counterfactually sensitive. For example identifying the ecological system that an echidna populations belongs to will include its predators, Goannas, and prey, Termites. If we claim that a local assemblage of populations belong to the same individual then those populations should map into a single ecological system. If this system is bound and has its own properties then it is a biological individual and an ecological community. If populations have a causal interaction profile which picks out the same ecological community with congruent boundaries and the same sub-parts, then that individual is robust.

The problem is that co-located populations often belong to radically different ecological systems. This is because causal relations in ecology are often asymmetrical and maps of organism distributions given by Global Information Systems (GIS) show that populations rarely spatially coincide[2]. Consider the factors relevant to a population of Spotted Quolls compared to their occasional prey, Greater Gliders. Individual Quolls roam over home ranges up to 3500 hectares moving between habitat fragments via wildlife corridors, while a Glider's home range is only 2 hectares and is locked within a local habitat fragment. Unless there is a very strong counterfactual dependence between these two populations the network of populations relevant to the Quolls will be radically different to the Gliders, as Quolls interact with populations that intersect with their large home ranges. Further, due to the radically different ranges and population densities there is a strong asymmetry between these populations. Differential changes in a local Glider population are unlikely to affect the Quoll population. Its range would include several Glider populations and they are generalist predators. But differential changes that increase the Quoll population would impact the

[2] See for example the Atlas of Living Australia for spatial distributions of populations across the continent.

Glider population as increased predation can have large impacts on small local populations. This creates an asymmetry; intervention on Gliders has little impact on Quolls but intervention on Quolls significantly impacts Gliders.

Population boundaries radically differ and the causal relations between populations are often asymmetrical. When these conditions are met, congruent boundaries are rare, and identifying the population network, and the space that network occupies, will be highly dependent on the initial choice of referent. Varying the starting population or property referred to in an assemblage will yield radically different descriptions of the ecological community. **Figure 2** shows the variation that can be displayed in a simple four population system for the parts and system outputs.

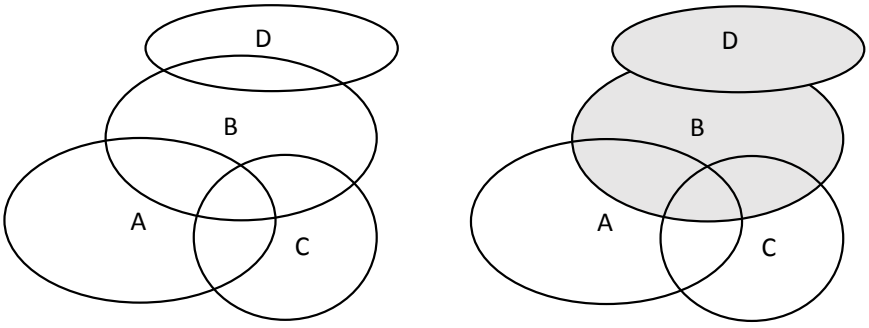


Figure 2. *Asymmetry and Congruence.* Each circle represents a population and its spatial range. These populations have the causal structure D → B → A, C→A. The relevant causal community of population A is all the local populations as it is causally influenced by all the other populations. If we want to find the relevant community for population B then it will include the shaded area as population B is only influenced by population D.

Ecology aims to explain how populations and their interactions result in system level properties like diversity, stability, or ecological services, e.g. water retention and biomass production. Local determinism supposes that stable relationships between persistent populations produce these properties; stable internal structure produces system level properties. Explanations of this type are *machine robust*; the system-level property is a result of a particular causal sequence of interactions between persistent parts. This is, however, a problematic assumption as populations are often highly transient. In one study of 100 biomes across earth, 75% of these systems had at least one in ten species disappear locally per decade (Dornelas, Gotelli et al. 2014). This is often coupled with little change in regional diversity as populations simply shift their distribution across the larger landscape (Thuiller, Slingsby et al. 2007). These studies add further evidence to those who believe that local ecological communities are often the wrong scale to look for law-like generalities in ecology (Ricklef, 1999; Lean and Sterelny, forthcoming). They claim that regional patterns better explain the local distribution and abundance of organisms than local patterns which are ephemeral and stochastic. These views explicitly reject the idea that local community identity is primarily maintained by internal composition.

Machine robust systems are relatively rare in ecology but there are ubiquitous higher level properties begging for explanation. One way we get these stable higher level outputs is via another type of robustness commonly found in complex systems, *ensemble robustness*. Ensemble robustness is when the system-level property is a product of diverse and varied parts filling the same functional role. The parts in the system do not have to be identical over time and space for the high level properties to be robust. For critical feedback loops for overall system functioning we often find huge redundancy; for example gum forest pollination is done by a range of evolutionary distinct actors including marsupial, insects, and birds. The statistical aggregation of the actions of local populations can stabilize ecological

output as a result of statistical averaging effects, biological insurance, and sampling effects (Bryant, 2010). Ecological systems also have outputs which are not just the simple aggregation of component population's actions. Diverse local species assemblages can have non-linear ecosystem outputs; for example, combinations of populations non-additively result in explosive combustion in forest fires (Van Altena et al, 2012). Due to all these factors higher level properties are ubiquitous in ecological systems even if there are no clear boundaries for these systems and the internal composition is unstable.

Despite the highly aggregational quality of ecological systems, ecological community properties are not uniformly a product of ensemble robustness; specific populations are sometimes necessary for ecological output. Keystone species, which have disproportionate impacts on assemblage composition, function like mechanisms with particular populations playing a necessary and causally specific role in maintaining whole system features. The importance of keystone species is controversial, with some ecologists pressing that there are not such strong relationships between single populations and assemblage features (Mills et al, 1993). But there is strong evidence that in some systems particular populations do play strong roles in regulating a cluster of populations in their assemblage (Ripple et al, 2001). Species can co-vary in tight relationships over geological periods that far outstretch local communities both spatially and temporally according to paleo-ecological evidence (Sterelny, 2001). Symbiotic relationships show similar tight co-variation between populations. This indicates that nested within larger assemblages we can find sets of populations with strong and persisting causal relations that are much more stable than the community as a whole.

To summarize, ecological communities are highly unsystematic systems, they lack clear boundaries and persistent internal identity, but they do have robust *parts* and robust *system outputs* via the variant aggregative interactions of their constituents. Any account of ecological community identity needs to be able to identify these explanatorily important

properties and fix the reference of the system that produces these properties. This is difficult as population networks will not in general be congruent over different choices of starting population as small changes in referent choice can result in a quite different network. But ecological communities are still causal systems. Indexical communities describe communities via the network of causal interactions between populations and provide a way to represent their causal structure.

4. Indexical Communities

On a first pass of the philosophy of ecology literature, accounts of ecological communities appear to split between treating populations as largely independent of each other, or describing them within an individuality framework. There are, however, other options which sit between these extremes with Sterelny proposing 'indexical communities' in contrast to ecological individuality (Sterelny, 2006). The following account of ecological communities supplements and develops indexical communities by providing the conceptual apparatus to identify robustness and utilizing the Woodwardian interventionist framework to fix the reference of the causal system involved (Woodward, 2005).

Simple indexical communities are ecological units which aim to describe the conditions that affect the demographics of single populations. Indexically described communities are one of the most useful and utilized ecological technique in conservation science. To preserve the critically endangered Hairy Nosed Wombat we need to know how much native grasses and tubers they eat, what is an unusual parasite load, how to separate them from wild dog populations, and competing gazers. These populations are indexed to the Wombat population as they have a causal impact on them. This framework has become commonplace due in part to conservation funding being directed to individual species preservation. The science then aims to find the conditions that lead to the preservation of a focal population.

These simple indexical communities are not thought to be very informative for community level properties as they are constructed with limited epistemic aims, i.e. explaining the influences on a single population. Due to the limited scope of such causal units they remain silent on certain, hopefully generalizable, community level features such as the relationship between diversity and stability (Sterelny, 2006). Further it is thought that information about one indexical community is difficult to apply to other assemblages due to the apparent limited nature of their scope. But we can rectify these problems by building into indexical communities the means for identifying machine robustness and ensemble robustness.

This is done by allowing for multiple decomposition of a local assemblage using Woodwardian Intervention. To be able to identify robustness in these ecological assemblage in their output or causal boundaries we need to have multiple starting points to investigate the unit. While indexical communities as they have previously been discussed are only around a focal population, this account of indexical communities expands the focal unit to a set of multiple populations. The stepwise procedure for identifying the relevant ecological community appears in **Box 1**, but here also is a description of the process. Take the starting set of populations and identify the indexical community of each individual population in the set. The indexical community for a population is identified by intervention: treat the focal indexical population as variable A. An alternative population variable, B, is said to be part of the same community, as well as a cause of A, if systematic intervention on B brings about change in A. Further, once we identify that a population variable has causal influence on the focal population we can ask whether intervention on populations that affect it also have 'downstream' effects on A. If so, then that population is also part of the community. Each population node introduced between the focal population and a population of interest will necessarily reduce the counterfactual relationship between them.

Box.1. *Indexical Communities can be built up from multiple indexical populations by the following procedure.*

i. Define the starting set of populations and/ or a community-level property (e.g. ecosystem output).

   a. If community-level property then identify the set of populations that contribute to the property.

ii. Identify the populations that are causally salient for the set of populations via intervention.

iii. Overlay the different networks of counterfactual dependencies from the specific populations.

iv. If multiple interventions pick out the same connection these are the robust relationships in a community.

This process yields a directed graphical map of the causal network indexed to population A. We repeat this procedure for all the populations in the starting set. The different causal maps are then compared. All the populations that causally contribute to the starting population are counted as part of the community. But the scope of the boundaries can be tweaked by varying the strength of the causal effect required for inclusion (Levins and Lewontin, 1985). By setting this parameter moderately high we avoid ecological holism, where each indexical community has a numerous nodes and as result each indexical community will overlap each other. Population network structures that appear from multiple different indexed populations are more robust. For example if there are populations that act like keystone species they will be part of all the directed graphs as they play necessary role in maintaining the population network structure.

What determines the starting set of populations? This is in part research interest defined but there are some obvious candidates. The first is including all the populations that cohabit

in a location. By identifying the network of populations that emanate out of co-habiting populations we can see to what extent this local ecological community is a causally cohesive unit. Alternatively we can look at community-level properties or outputs by starting with the set of populations that are thought to contribute to this community-level feature, for instance water filtration around a lake. The ecological structure that yields this output (filtration) is of economic interest and indexical communities identify the populations that need to be preserved to maintain this community output.

What I find the most exciting aspect of this framework is it gives us the means to preserve two different important conservation units that have previously been criminally under-described and referentially underdetermined. These are phenomenological communities and biodiverse communities. Phenomological communities are the communities that the folk who are interested in and spend time in the environment perceive. Environmentalists and the public often have an interest in preserving particular assemblages that are familiar from their experience of the wild. These assemblages include charismatic mammals, audible bird-life, visually stimulating angiosperms, and imposing trees. To fix the reference of such local assemblages we include in the starting set the phenomenologically prominent populations in a local area. For example if you want to find the community of a Blue Gum forest you include Blue Gums, lyrebirds, and Waratahs and identify the populations relevant to them.

The second conservation-based community is a biodiverse community. The preservation of biodiversity has been the primary goal of conservation science for the last 30 years but 'biodiversity' is ill-defined. Two major philosophical positions regarding biodiversity are conventionalism - biodiversity is the features of biological difference that community stake-holders value - or realism - there are privileged carvings of biological difference which we should value (Sarkar, 2006; Maclaurin and Sterelny, 2008; Lean and Maclaurin, 2016). For

either position we can identify the populations which represent *biodiversity* in that particular local area and then use this procedure to find the relevant larger ecological community.

By allowing the starting set to be determined by the interested parties we are able to tailor the indexical community to fulfil both the epistemic and normative roles that community ecology and conservation science requires.

5. Upshots of Indexical Communities

Built into this methodology is the means of assessing an ecological community in several ways. The first is the invariance and production of community level properties. To explain how the starting assemblage produces a particular community-level property, be it stability of population network structure or an ecosystem output like fire likelihood, we need to identify the counterfactual interventions that affect that property. We do this so we can assess the invariance of the populations and their relationship to these properties. If particular populations appear in multiple networks in the same sequence, those parts of the system are robust, so we can gain a picture of the way these stable causal relations yield community properties. The indexical community identifies the descriptively robust features of the system under inquiry. Machine robust parts of the ecological network will always be descriptively robust. Weak aggregational interactions also bring about community-level phenomena. These are instantiated by many pathways, which have modest strength. To gain a sense of the relationship between the aggregative system and these properties we need to fix the identity of the system in question. Indexical communities provides a precise way to refer to such weak 'systems' and in by doing so provides a guide for further research into the relations between populations and community-level properties.

For the ontological question of whether communities are real, indexical communities provides an answer. If the same causal structure appears from multiple starts and has robust

boundaries then we have a robust ecological community. It is, however, more likely that we will find that we have only partial overlap between the causal maps. This acts to identify the descriptive robust sub-systems within the community. As a result, this framework provides a more fine-grained and specific way of identifying whether a particular local ecological community is a system that acts like an individual, like an organism, or an aggregate, like gas particles in a beaker. If there is no causal connections between the starting populations then this is not a unitary community. So this methodology acts not just as a descriptive tool but also an existence test. Depending on referent choice, there can be multiple precisifications of a unitary community or none.

By describing communities using a causal graph network description, we open them up to a range of formal methods of assessment. Modularity of the system and the sub-systems is one important feature. Modular clusters of causal interactors make a system more bounded and can account for particular system outputs. Formal methods like the Girvan-Newman algorithm can quantify such structures identifying modular grouping and boundaries in complex systems (Givarn and Newman, 2002).

This is all to say we can assess an indexical ecological community in terms of the invariance of its system properties, its modularity, and its descriptive robustness. If an indexical community is completely modular, descriptively robust, and has highly invariant system properties then it will be a biological individual. But communities so rarely satisfy these conditions that we need an alternative framework. There is more to biology than just the study of individuals and this proposal gives an alternative framework to describe complex biological systems.

References

Bryant, Rachael. 2012. “What If Ecological Communities Are Not Wholes?” *The Environment: Philosophy, Science, and Ethics*, ed. William Kabasenche, Michael Oourke, and Matthew Slater, MIT Press: 37-56.

Dornelas, Maria., Nicholas Gotelli, Brian McGill, Hideyasi Shimadzu, Faye Moyes, Caye Sievers, and Anne Magurran 2014. "Assemblage time series reveal biodiversity change but not systematic loss." *Science* 344 (6181): 296-299.

Girvan, Michelle., and Mark Newman. 2002. “Community structure in social and biological networks.” *Proceedings of the national academy of sciences*, *99*(12): 7821-7826.

Lean, Christopher., and James Maclaurin. 2016. “The Value of Phylogenetic Diversity” In *Biodiversity Conservation and Phylogenetic Systematics*. ed. Roseli Pellens and Philippe Grandcolas, Springer Press.

Lean, Christopher., and Kim Sterelny. Forthcoming. “Biodiversity and Ecological Hierarchy.” In *The Routledge Handbook of the Philosophy of Biodiversity.* ed. Justin Garson, Anya Plutynski, and Sahotra Sarkar

Levins, Richard., and Richard Lewontin. 1985. *The Dialectical Biologist*. Cambridge: Harvard University Press.

Maclaurin, James., and Kim Sterelny. 2008. *What is Biodiversity?* Chicago: University of Chicago Press.

Mills, Scott, Michale Soulé, and Daniel Doak. 1993. “The keystone-species concept in ecology and conservation.” *BioScience*, *43*(4): 219-224.

Odenbaugh, Jay. 2007. “Seeing the Forest *and* the Trees: Realism about Communities and Ecosystems.” *Philosophy of Science*, 74(5): 628–641.

Odenbaugh, Jay. Forthcoming. “Conservation Biology.” *The Stanford Encyclopaedia of Philosophy* Ed: E. Zalta

Queller, David. 2000. “Relatedness and the fraternal major transitions.” *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 355(1403): 1647-1655.

Ricklefs, Robert. 2008. "Disintergration of the Ecological Community." *American Naturalist* 272(6): 741-750.

Ripple, William., Eric Larsen, Roy Renkin, and Douglas Smith. 2001. “Trophic cascades among wolves, elk and aspen on Yellowstone National Park’s northern range.” *Biological Conservation*, *102*(3): 227-234.

Sarkar, Sahotra. 2005. *Biodiversity and Environmental Philosophy*. Cambridge: Cambridge University Press.

Sterelny, Kim. 2001. The reality of ecological assemblages: A palaeo-ecological puzzle. *Biology and Philosophy*, *16*(4): 437-461.

Sterelny, Kim. 2006. "Local Ecological Communities." *Philosophy of Science* 73(2): 215-231.

Thuiller, Wilfried., Jasper Slingsby, Sean Privett, and Richard Cowling. (2007). "Stochastic species turnover and stable coexistence in a species-rich, fire-prone plant community." *Public Library Of Science One* 2(9): 938.

Van Altena, Cassandra., Richard van Logtestijn, William Cornwell, and Johannes Cornelissen. 2012. Species Composition and Fire: Non-Additive Mixture Effects on Ground Fuel Flammability. *Frontiers in plant science*. 3:63.

Wimsatt, William. 2007. "Complexity and Organisation." *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge: Harvard University Press, 179-192.

Woodward, James. 2005. *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

The deflationary view of information reloaded: communication and manipulability

Abstract

Timpson's deflationary view of information is an innovative and well articulated view that had a great impact on the philosophy of physics community. However, recently some of the arguments supporting the deflationist view have been critically reviewed. The aim of this paper is to retain the general idea behind Timpson's proposal, but replacing the conflictive elements used to support his thesis with new argumentative resources based on the notion of manipulability.

1. Introduction

The central role played by the notion of information in contemporary science led to a significant growth of the literature on the subject. In the field of the philosophy of science, Christopher Timpson's works stand out for their soundness and wide scope: the domain of application of Shannon's theory (Timpson 2003), the relation between information transmission and quantum entanglement (Timpson 2005), the interpretation of teleportation (Timpson 2006), the relation of quantum information with the interpretations of quantum mechanics (Timpson 2008, 2013), among others. In this paper we will only focus on his deflationary view of technical information, according to which the term 'information' is an abstract noun and, as a consequence, information is not a concrete physical substance.

This innovative and well-articulated view had a great impact on the philosophy of physics community, becoming a kind of orthodox view about the concept of information. However, recently some of the arguments supporting the deflationist view have been critically reviewed (Lombardi, Fortin and López 2016). The aim of this paper is to retain the general idea behind Timpson's proposal, but replacing the conflictive elements used to support his thesis with new argumentative resources based on the notion of manipulability.

2. Timpson's Deflationary View and Its Difficulties

Although Timpson's deflationary proposal has many facets, its main goal is to eradicate the view of information as something material, a substance or stuff (Timpson 2004, 20; 2008, 28; 2013, 34-36): "one should not understand the transmission of information on the model of transporting potatoes, or butter, say, or piping water." (Timpson 2008, 31). Besides its plausibility, this perspective dissolves the problems related to with communication based on entanglement. In particular, Timpson (2006) cuts the Gordian knot of teleportation: if

'information' is an abstract noun, the question about how information "travels" from the source to the destination in teleportation is dissolved.

Timpson's argumentative strategy begins by introducing the difference between "bits" and "pieces" of information (2008. 'Quantity' replaces 'bits' in 2013, 16), in order to show that in both cases information is an abstract item. The notion of bits of information refers to the amount of information that a source of information produces. A piece of information is "*what* the output of a source (quantum or classical) is" (2008, 27; emphasis in the original). On this basis, the argument for the abstractness of information runs as follows. On the one hand, information qua-quantity is abstract because quantities are abstract. On the other hand, the abstractness of information qua-piece relies on the philosophical distinction between *types* and *tokens*: "one should distinguish between the concrete systems that the source outputs and the type that this output instantiates." (Timpson 2004, 22). The piece of information is not the token produced at the source, but the corresponding type; and since types are abstract, then information qua-piece is abstract.

The argumentative strategy sounds very appealing, and perhaps this fact explains the high impact of Timpson's deflationary view. However, when subjected to further scrutiny, certain difficulties come to light.

According to Timpson, when the source of information produces a message, what we want to transmit is not the sequence of the states themselves: "one should distinguish between the concrete systems that the source outputs and the type that this output instantiates." (Timpson 2004, 22). The goal of communication is to reproduce at the destination another token of the same type: "What will be required at the end of the communication protocol is either that another token of this type actually be produced at a distant point (as a consequence of the production of the initial token); or at least that it be *possible* to produce it there (as a

consequence of the initial production) by a standard procedure." (Timpson 2013, 23, emphasis in the original; see also Timpson 2008, 25).

The problem here is that the goal of communication is not to reproduce at the destination a token of the same type as that produced at the source. As Shannon stresses, "[t]he significant aspect is that the actual message is one *selected from a set* of possible messages." (1948, 379, emphasis in the original). This means that the goal of communication consists in *identifying at the destination the state produced at the source*. Therefore, the success criterion is given by a one-to-one or a one-to-many mapping from the set of letters of the source to the set of letters of the destination. As Duwell (2008, 200) correctly points out, this mapping is arbitrary; then, the states of the source and the states of the destination may be of a completely different nature: the source may be a dice and the destination a dash of lights; or the source may be a device that produces words in English and the destination a device that operates a machine. It is difficult to understand in what sense a face of a dice and a light in a dash are tokens of a same type (see full argument in Lombardi, Fortin and López 2016).

Perhaps with the purpose of stressing the arbitrariness of the success-defining mapping, in his 2013 book Timpson generalizes the type-token distinction in terms of *sameness of pattern or structure*: "the distinction may be generalized. The basic idea is of a pattern or structure: something which can be repeatedly realized in different instances" (2013, 18). However, this new move is not free of difficulties. First, sameness of structure is a purely formal relation, which cannot be simply identified with the philosophical relation between tokens of the same type: a type needs to have some content to be able to identify its tokens. In other words, the distinction between types and tokens is not merely formal or syntactic: being tokens or a same type is not an arbitrary relation. Admitting arbitrary functions as defining the relation of being tokens of the same type leads to admit that any two things arbitrarily chosen

can always be conceived as tokens of the same type and, thus, trivializes the distinction type-token (see Wetzel 2014).

However, the main difficulty is technical (Lombardi, Fortin and López 2016). The characterization of the goal of communication in terms of sameness of structure disregards the possibility of noisy situations. In the presence of noise, the states of the source and the states of the destination are linked by a one-to-many mapping; nevertheless, this does not prevent a successful communication, since the states of the source can still be identified by means of the states of the destination. These noisy cases are precisely the situations of real interest in the practice of communication engineering.

These objections do not aim at undermining Timpson's proposal. On the contrary, the challenge is to try to preserve his deflationist view of information, but avoiding those criticisms. This will be the task of the following sections.

3. Information and Communication

If the deflationary strategy is developed in the line followed by Timpson, the goal of rejecting the stuff-view of information is reached but at the price of turning information into a purely formal concept. This is the trend of some recent textbooks, which introduce information theory in an exclusively formal way, with no mention of sources, destinations or signals: the basic concepts are defined in terms of random variables and probability distributions over their possible values. It is only when the formalism has been presented that the theory is applied to the traditional case of communication (see the extensively used book of Cover and Thomas 1991). From this perspective, information theory is a new chapter of the theory of probability: the word 'information' does not belong to the language of empirical sciences, it has no extralinguistic reference in itself. Its "meaning" has only a syntactic dimension. As a consequence, the generality of the concept of Shannon information derives from its

exclusively formal nature; this generality is what makes it a powerful formal tool for empirical science, applicable to a variety of fields. (Lombardi, Fortin and Vanni 2015).

Although this purely formal conception of information has its advantages, this is not the conceptual framework of Timpson's argumentation. On the contrary, his discussion is framed within a particular context, in which information is much more than mere correlation. In fact, it is important to recall that, even in the domain of statistical information (as different from semantic information, see Floridi 2015), different contexts can be distinguished, each one with its particular formal resources to deal with specific goals.

In the *computational context*, information is something that has to be computed and stored in an efficient way. The algorithmic or Kolmogorov complexity measures the minimum resources needed to effectively reconstruct an individual message (Solomonoff 1964; Kolmogorov 1965; Chaitin 1966): it supplies a measure of information for individual objects taken in themselves, independently of the source that produces them. In this context, the basic question is the ultimate compression of individual messages; the main idea is that the description of some messages can be compressed considerably if they exhibit enough regularity. The Kolmogorov complexity of a message is, then, defined as the length of the shortest possible program that produces it in a Turing machine. In the *communicational context*, whose classical formalism is Shannon theory (Shannon 1948, Shannon and Weaver 1949), information is primarily something that has to be transmitted between to points for communication purposes. The formalism was designed to solve certain specific technological problems in communication engineering, in particular, to optimize the transmission of information by means of physical signals whose energy and bandwidth is constrained by technological and economic limitations. Although the computational and the communicational contexts are the traditional ones, they are not the only ones. For instance, in an *inferential context* the interest is to find a universally good prediction procedure on the basis of the

possessed data, where "good" involves a version of Occam's Razor: 'The simplest explanation is best.' The *thermodynamic context* is devoted to relate information to entropy and to explain the entropy increase in terms of informational concepts and arguments. In a *gambling context* the problem is to use informational resources to formalize a gambling game, by representing the wealth at the end of the game as a random variable and the gambler as a subject that tries to maximize that variable.

Timpson is clearly thinking in communication which, although independent from any informational content –"The semantic aspects of communication are irrelevant to the engineering problem." (Shannon 1948, 379)–, is not a merely syntactic notion. Many different definitions of the concept of communication can be found in the literature, most of them involving semantic notions such as meaning, epistemic notions such as understanding, and even notions referring feelings and emotions. Here we are not interested in giving a definition of communication, but only in isolating some of its essential notes, in order to identify in which situations it can be said that there is communication and, consequently, transmission of information.

Regardless of how the quantity of information is defined, there are certain minimum elements that can be abstracted to characterize a communicational context. From a very general perspective, communication requires a source S with its different states, which produces the information to be transmitted, a destination D with its own states, which receives the information, and a channel through which information is transmitted from the source to the destination. In the context of this abstract framework, communication requires that a certain action performed at the source modifies the destination so as to establish a correlation between the state of the source and the state of the destination. Let us notice that this characterization does not involve knowledge, not even in the weak sense of a subject that identifies what state occurred at the source from knowing the state occurred at the destination: the state of the

destination can be manipulated from the source for exclusively control purposes. Moreover, perfect correlation is not required: non-perfect correlation, manifested as equivocity and/or noise, can be corrected my means of redundancy and/or filters that preserve communication.

This characterization, although very abstract, includes two essential notes, which are not independent but linked to each other:

- *Asymmetry*. Communication is an asymmetric process: the source sends information and the destination receives it. Although in a following stage the roles can be interchanged, in each run of the communication process source and destination are clearly different and cannot be confused.
- *Production*. What happens at the source modifies what happens at the destination, produces a specific change of the state of the destination. In other words, the asymmetry is not a merely formal relationship, but a physical connection that links events occurred in different space-time locations.

These two features are not manifest in Shannon's formalism taken at face value, without adding explanations related to producing, sending and receiving information (see Cover and Thomas 1991). They can neither be obtained from Timpson's deflationist view of information: asymmetry and production find no place in the context of a view that conceives the link between source and destination merely as sameness of form. Therefore, an additional ingredient is necessary to identify situations of transmission of information in a communicational context.

If, independently of the nature of source, channel and destination, communication requires that what happens at the source in a certain way produces what happens at the destination, perhaps the notion suitable in this case is that of *causation*. Moreover, the causal connection between source and destination would recover the asymmetry of the

communication process. Although this seems a promising strategy, the main challenge that it has to face is the elucidation of the very concept of causation.

4. Communication and Manipulability

Although the appeal to causation may offer a clue to elucidate the concept of information in the communicational context, the risk is to try to elucidate an obscure concept by means of another even more obscure: the notion of causation is one of the most controversial topics in the history of philosophy.

The traditional approaches to causation are usually classified into two categories: the counterfactual approaches and the physical approaches. The first ones carry the burden of finding a proper semantics for counterfactuals. The physical approaches seem to be more adequate in a physical context of communication. From the physical perspective, causation has been conceived in terms of energy flow (Fair 1979), of physical processes (Dowe 1992), and of property transference (Kistler 1998): all these views involve physical signals or space-time connections between cause and effect. For this reason, the physical approaches to causation might have been particularly useful to elucidate the concept of information in the traditional cases of communication, as those that constituted the original field of application of Shannon theory. Those cases, the most common in engineering, are constrained by the well-known *dictum* 'no information without representation': the transmission of information between two points of the physical space necessarily requires an information-bearing signal, that is, a physical process propagating from one point to the other. As expressed by Landauer (1996, 188), "[i]nformation is not a disembodied abstract entity; it is always tied to a physical representation. It is represented by engraving on a stone tablet, a spin, a charge, a hole in a punched card, a mark on a paper, or some other equivalent."

However, at present the landscape has drastically changed with the advent of quantum information theory. In fact, teleportation challenges the traditional assumption about the inescapable need of a physical signal carrying the information through space. Broadly speaking, an unknown quantum state $|\chi\rangle$ is transferred from Alice to Bob with the assistance of a shared pair of particles prepared in an entangled state and of two classical bits sent from Alice to Bob. The perplexity of the case lies in that –among other facts– the information is transferred from Alice to Bob without any physical signal other than the classical channel through which the classical bits are transmitted. As a consequence, a conception of causation based on physical interaction through space and time cannot account for the transmission of information in the case of the paradigmatic example of entanglement-assisted communication.

On the other hand, independently of any philosophical discussion, both in the everyday life as in science people act as if there were real causal links, without considering whether or not there is a space-time connection between the cause and its effect. Regarding causes, anyone distinguishes the case of pain due to burn injury from the appearance of the paperboy when the sun rises. Similarly, a chemist clearly distinguishes the causal action of a catalyst in increasing the rate of a reaction from the mere correlation between the melting point and the color of an element.

The *manipulability accounts of causation* intend to capture this intuitive distinction. Their basic idea is that it is possible to draw the distinction between cause-effect relationships and mere correlations by means of the notions of manipulation and control. As Cartwright (1979) stresses, causal relationships ground the distinction between effective and ineffective strategies: an effective strategy proceeds by intervening at a cause in order to obtain a desired outcome. In other words, only causal relationships, but not mere correlations, are exploitable by us in order to bring about a certain outcome (Frisch 2014).

There are different manipulability accounts of causation. According to the early versions, causal terms need to be reduced to non-causal terms, such as free agency (von Wright 1971; Price 1991; Menzies and Price 1993). These first manipulability versions received several criticisms. On the one hand, they were charged of circularity: since "doing" and "producing" are already causal notions, they cannot be legitimately used to define the notion of causation. On the other hand, manipulability is an anthropocentric notion; then, the resulting concept of causation is not sufficiently general; for instance, it is not be able to identify the relationship between the gravitational attraction of the moon and the motion of the tides on the earth as a causal relation.

The interventionist version of the manipulability account of causation (Woodward 2003, 2007; Hausman and Woodward 1999; Pearl 2000) comes to solve those criticisms. Given the variables *X* and *Y*, "the intuitive idea is that an intervention on *X* with respect to *Y* changes the value of *X* in such a way that if any change occurs in *Y*, it occurs only as a result of change in the value of *X* and not from some other source" (Woodward 2003, 14). In this case, it can be said that the relationship between *X* and *Y* is a genuine case of causation. According to Woodward, the circularity criticism does not apply because the interventionist approach does not intend to define causation in terms of non-causal notions, but to delimitate the domain of causation by means of the possibility of control or manipulation: the response to interventions is used as a probe to know whether or not a certain relation is causal (Woodward 2003, 21). As Frisch (2014, 78) puts it, "the results of interventions into a system are a guide to the causal structure exhibited by the system." On the other hand, the interventionist faces the charge of anthropocentrism by arguing that the concept of intervention must be understood without reference to human action. According to Woodward, the consideration of *possible* interventions admits a counterfactual formulation, which makes sense of causal claims in situations where interventions do not occur and even in cases in which they are impossible in

practice. Nevertheless, the interventionist approach is not a counterfactual view of causation because counterfactuals are not applied to the very relationship whose causal nature is to be determined.

Besides the traditional charges of circularity and anthropocentrism, other criticisms have been directed toward the interventionist approach to causation (see Woodward 2013). One of them is related to the use of counterfactuals: since the truth conditions for counterfactuals can be explained in terms of laws, the appeal to interventions is not necessary (Hiddleston 2005). In turn, Cartwright characterizes the interventionist approach as "operationalist": it admits a single criterion to test causation, and leads to "withhold the concept [of cause] from situations that seem the same in all other aspects relevant to its application just because our test cannot be applied in those situations" (Cartwright 2002, 422).

Independently of how appropriate these criticisms are, they are directed toward a position that intends to supply the elucidation of the very concept of causation. Here we will not enter into the debate about whether the interventionist approach reaches its goal or not, because our concerns about causation are more modest. We are not interested in supplying a characterization of causation applicable in every circumstance in which the causal talk makes sense. Our only aim here is to explore the possibility of appealing to interventionist causation to characterize the informational relation between source and destination in a communicational context.

In this context, the charge of anthropocentrism is innocuous: here we are not interested in the moon causing tides or in the motion of tectonic plates causing earthquakes. Our interest is confined to cases of communication, in which there is always a deliberate intervention on the source of information with the purpose to change the state of the destination. Moreover, Cartwright's worries are beyond our limited scope: the fact that the interventionist concept of cause cannot be applied in certain relevant causal situations is not a problem if those situations

do not involve communication. In our case, causation is used only as a probe tool to know whether there is transmission of information or not in the communicational context independently of signals and interactions between source and destination.

Regarding the objection to the use of counterfactuals, it does not apply in our context of interest. In fact, counterfactuals are introduced in the interventionist approach to deal with cases where the intervention on the cause is physically or practically impossible. But in situations of transmission of information the interventions on the source are always physically and practically possible. Even more, since the messages to be transmitted are embodied in sequences of the states of the source, the possibility of controlling the state of the source is an essential requirement for communication: the nature of communication itself includes that possibility; it makes no sense to conceive a source of information whose states cannot be modified.

Despite we found this strategy to elucidate the notion of information in communication completely reasonable, those who outright reject the interventionist account of causation might transitively reject the strategy. Therefore, perhaps it is convenient to slightly modify the strategy in the following sense. Up to now we searched for the ingredient that would account for the feature of production that distinguishes communication from mere correlation, and found it in causation, in particular, in its interventionist version. But we can reach the same goal without appealing to the concept of causation, and by using the interventionist view of causation as a mere inspiration for designing a *manipulability approach to information in the communicational context*. From this perspective, causation does not matter: the interventionist arguments serve the function of delimitating the domain of communication by means of the possibility of control or manipulation. To put it in another way, the response to interventions is used as a probe to know whether there is transfer of information or not.

This manipulability approach to information applies successfully to entanglement-assisted communication, where there is no information-carrier signal other than that carries the classical bits. Although teleportation is based on EPR correlations, it is not a mere EPR-experiment. In fact, Alice not only counts with a particle of the entangled pair, but she also has access to the state $|\chi\rangle$ to be teleported, and to the two two-state classical systems needed to send the two bits of information through the classical channel. Since communication requires those three elements, the intervention does not need to act on the entangled pair, but it can operate on the other two elements. For instance, the intervention on Alice's end may change the state to be teleported, from $|\chi\rangle$ to $|\varphi\rangle$: as a consequence, something changes in Bob's end, since he will recover $|\varphi\rangle$ and not $|\chi\rangle$. Or the intervention might block one of the classical systems that Alice sends to Bob: in this case, Bob would be unable to recover the teleported state. It is worth stressing that we can be sure about the consequences of these interventions independently of whether the entangled pair is interpreted as consisting of two systems or as a single holistic whole.

5. Conclusions

Traditional engineers could feel comfortable conceiving information as something that is transferred from source to destination through space: in fact, before the advent of quantum information theory, the need of a signal acting as a carrier of information was a basic and indisputable assumption in the training of communication engineers. But entanglement-assisted communication changed the panorama and raised new challenges for the concept of information. In this context, Timpson proposed a deflationist view of information that overcame the obstacles posed by the new forms of communication. However, his strategy to support deflation, inspired in resources coming from the philosophy of language, show some weak spots when considered from a broader perspective.

In this paper we intended to follow the deflationist trend opened by Timpson, but avoiding the difficulties derived from his particular line of argumentation. With this purpose, we seek for a foundation of the central features of communication –asymmetry and production– independent from travelling carrier signals. The inspiration came from the manipulability accounts of causation. In particular, we used the strategy of the interventionist view adapted to the case of communication to identify the cases in which there is information transference. Of course, this is only a first step of a research that deserves to be further developed. Nevertheless, this deflationist perspective not only can be successfully applied to account for teleportation; it also seems to be a perspective more natural for those scientists and technologists interested in the practical exploitation of quantum entanglement for communication.

6. References

Cartwright, Nancy. 1979. “Causal Laws and Effective Strategies.” *Noüs* 13: 419-37.

–––. 2002. “Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward.” *The British Journal for the Philosophy of Science* 53: 411-53.

Chaitin, Gregory. 1966. “On the Length of Programs for Computing Binary Sequences.” *Journal of the Association for Computing Machinery* 13: 547-69.

Cover, Thomas, and Joy Thomas 1991. *Elements of Information Theory*. New York: JohnWiley & Sons.

Dowe, Phil. 1992. “Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory.” *Philosophy of Science* 59: 195-216.

Duwell, Armond. 2008. “Quantum Information Does Exist.” *Studies in History and Philosophy of Modern Physics* 39: 195-216.

Fair, David. 1979. “Causation and the Flow of Energy.” *Erkenntnis* 14: 219-50.

Floridi, Luciano. 2015. “Semantic Conceptions of Information.” In *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), ed. Edward N. Zalta, URL = <http://plato.stanford.edu/archives/spr2015/entries/information-semantic/>.

Frisch, Mathias. 2014. *Causal Reasoning in Physics*. Cambridge: Cambridge University Press.

Hausman, Daniel, and James Woodward. 1999. “Independence, Invariance, and the Causal Markov Condition.” *The British Journal for the Philosophy of Science* 50: 521-83.

Hiddleston, Eric. 2005. “Review of *Making Things Happen*.” *Philosophical Review* 114: 545-47.

Kistler, Max. 1998. “Reducing Causality to Transmission.” *Erkenntnis* 48: 1-24.

Kolmogorov, Andréi. 1965. “Three Approaches to the Quantitative Definition of Information.” *Problems of Information Transmission* 1: 4-7.

Landauer, Rolf. 1996. “The Physical Nature of Information.” *Physics Letters A* 217: 188-93.

Lombardi, Olimpia, Sebastian Fortin, and Cristian López. 2016. “Deflating the Deflationary View of Information.” *European Journal for Philosophy of Science*, on line first.

Lombardi, Olimpia, Sebastian Fortin, and Leonardo Vanni. 2015. “A Pluralist View about Information.” *Philosophy of Science* 82: 1248-59.

Menzies, Peter, and Huw Price. 1993. “Causation as a Secondary Quality.” *The British Journal for the Philosophy of Science* 44: 187-203.

Pearl, Judea. 2000. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Price, Huw. 1991. “Agency and Probabilistic Causality.” *The British Journal for the Philosophy of Science* 42: 157-76.

Shannon, Claude. 1948. “The Mathematical Theory of Communication.” *Bell System Technical Journal* 27: 379-423.

Shannon, Claude, and Warren Weaver. 1949. *The Mathematical Theory of Communication*. Urbana and Chicago: University of Illinois Press.

Solomonoff, Ray. 1964. “A Formal Theory of Inductive Inference.” *Information and Control* 7: 224-54.

Timpson, Christopher. 2003. “On a Supposed Conceptual Inadequacy of the Shannon Information in Quantum Mechanics.” *Studies in History and Philosophy of Modern Physics* 34: 441-68.

–––. 2004. *Quantum Information Theory and the Foundations of Quantum Mechanics*. PhD diss., University of Oxford (arXiv:quant-ph/0412063).

–––. 2005. “Nonlocality and Information Flow: The Approach of Deutsch and Hayden.” *Foundations of Physics* 35: 313-43.

———. 2006. "The Grammar of Teleportation." *The British Journal for the Philosophy of Science* 57: 587-621.

———. 2008. "Philosophical Aspects of Quantum Information Theory." In *The Ashgate Companion to the New Philosophy of Physics*, ed. Dean Rickles, 197-261. Aldershot: Ashgate Publishing. Page numbers taken from arXiv:quant-ph/0611187.

———. 2013. *Quantum Information Theory and the Foundations of Quantum Mechanics*. Oxford: Oxford University Press.

Von Wright, Georg. 1971. *Explanation and Understanding*. Ithaca: Cornell University Press.

Wetzel, Linda. 2014. "Types and Tokens." In *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), ed. Edward N. Zalta, URL = <http://plato.stanford.edu/archives/spr2014/entries/types-tokens/>.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

———. 2007. "Causation with a Human Face." In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, eds. Huw Price and Richard Corry, 66-105. Oxford: Oxford University Press.

———. 2013. "Causation and Manipulability." In *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition), ed. Edward N. Zalta, URL = <http://plato.stanford.edu/archives/win2013/entries/causation-mani/>

Word count: 4978

Science and proven experience: a Swedish variety of evidence-based medicine?

Johannes Persson, Niklas Vareman, Annika Wallin, Lena Wahlberg, and Nils-Eric Sahlin

Word count: 4849

Abstract:

A key question for evidence-based medicine (EBM) is how best to model the way in which EBM should "[integrate] individual clinical expertise and the best external evidence" (Sackett et al. 1996). We argue that the formulations and models available in the literature today are modest variations on a common theme and face very similar problems. For example, both the early and updated models of evidence-based clinical decisions presented in Haynes, Devereaux and Guyatt (2002) assume (with Sackett, et. al., 1996) that EBM consists of, among other things, evidence from clinical research and clinical expertise. On this *A-view*, EBM describes all that goes on in a specific *justifiable* medical decision. There is, however, an alternative interpretation of EBM, the *B-view*, in which EBM describes just one component of the decision situation (a component usually based on evidence from clinical research) and in which, together with other types of evidence, EBM leads to a justifiable clincial decision but does not describe the decision itself. This B-view is inspired by a 100-years older version of EBM, a Swedish standard requiring medical decision-making and practice to be in accordance with 'science and proven experience'. In the paper we outline how the Swedish concept leads to an improved understanding of the way in which scientific evidence and clinical experience *can* and *cannot* be integrated in

light of EBM. In addition the paper sketches the as yet unexplored historical background to EBM.

1. Introduction

> EBM is actually only a reformulation of the motto 'science and proven experience' (Werkö et al. 2002, 3478, our translation)

Globally, evidence-based medicine (EBM) is favoured in the public sector. In Sweden a bipartite, partly overlapping standard that is more than 100 years older than EBM applies in the public sector and parts of the private sector as well. This Swedish standard requires decision-making and practice to be based on both science and proven experience, or *vetenskap och beprövad erfarenhet* (VBE), and indeed leading Swedish physicians often think of EBM as a reformulation of the Swedish standard. Lars Werkö, "the icon beyond all comparison in Swedish health care" (Hont 2009) is a clear example (e.g. see Werkö et al. 2002 above). Since its first legal application in Swedish health care in the late 1800s, the VBE standard has been pressed into service in law and public policy in areas as diverse as medicine and health care, education, environmental risk assessment, veterinary care and social work. We shall focus on the application of VBE in medicine (VBE-M).

The Swedish concept of VBE-M helps us to understand the ways in which scientific evidence and clinical experience both *can* and *cannot* be integrated within EBM. The similarities and dissimilarities between VBE-M and EBM shed light on the capacity of

EBM to "[integrate] individual clinical expertise and the best external evidence" (Sackett et al. 1996). This most influential and elusive ambition of EBM (Sackett et al. 1996) is the primary focus of the current paper. In addition, however, VBE-M helps to bring out the historical background to EBM.

2. EBM and VBE-M: Similarities and dissimilarities

Prima facie it makes sense to compare EBM and VBE-M. (1) Both introduce evidentiary standards for decision-making (not a standard for science as such). (2) Each promotes the goal of making more use of science and sound evidence in practical decision-making. But (3) the two approaches diverge, on the surface at least, when it comes to the types of evidence that should be allowed to influence practical decision-making.

(1) EBM highlights the decision-making context. It is primarily about the professions and the connection between the academic disciplines and the professions; only by implication is it about the disciplines themselves. Several of the leading articles on EBM were published in the *British Medical Journal*, which has the slogan: *helping doctors make better decisions*. Rosenberg's and Donald's well-known 1995 *BMJ* paper is entitled "Evidence based medicine: an approach to clinical problem-solving".

Similarly, in Swedish law VBE is the gold standard for decision-making and practice, especially in health care (VBE-M). For example, VBE-M states that medical practice must be based on science and proven experience. Health care workers who do *not* provide care

in accordance with VBE-M can be criticized by the Health and Social Care Inspectorate and even be held responsible according to penal law. VBE also helps to define patients' rights to reimbursement for expenses associated with treatments in other European countries. The legislative use of science and proven experience illustrates that the notion is intended for policy-making and practical decision-making.

(2) To a large extent EBM focuses on the need to make more, and better, use of research findings in clinical decision-making:

> Evidence based medicine is the process of systematically finding, appraising, and using contemporaneous research findings as the basis for clinical decisions. For decades people have been aware of the gaps between research evidence and clinical practice, and the consequences in terms of expensive, ineffective, or even harmful decision making. Inexpensive electronic databases and widespread computer literacy now give doctors access to enormous amounts of data. Evidence based medicine is about asking questions, finding and appraising the relevant data, and harnessing that information for everyday clinical practice. (Rosenberg and Donald 1995)

The focus, claims (Eddy 2005, 14) , "is on educating physicians to help them bring more research and evidence into their individual decisions about individual patients." The same goes for VBE-M. However, before we proceed it should be noted that, in health care at

least, the Swedish notion is not explicitly defined in any official documents. One will search in vain for suitable stipulations to guide applications of the laws in which the expression occurs. Hence its characteristics have to be inferred from its many applications, i.e. its use and history. We will therefore present a sketch of the history of VBE-M (and to a lesser extent EBM).

It was probably no accident that the Swedish concept emerged in health care regulations in the late 1800s (the concept itself is, at least, somewhat older; almost exactly the same formulation occurs in the oath that were taken by those who were awarded the licentiate degree in medicine in Uppsala, Lund, and Stockholm from 1829 and onwards). It is sometimes argued that the mid-1800s were dominated by lack of confidence in the therapeutic methods available in Sweden and elsewhere. Scholarly work referring to this period uses terms such as “doubt disease” (Fåhræus 1950, 98), “therapeutic nihilism” (Danek 1969, 65), “the bankruptcy of therapy”, and “crisis in medical self-confidence” (Stolt 1994, Chapters 4 and 5). Mid-1800s advances in basic medical science did not reach practitioners and were generally of little therapeutic consequence (Porter 1995), and this was especially so, perhaps, for the typical Swedish countryside doctor:

> Practitioners in the countryside used trial and error, and as late as 1850 they had little use of medical science in their everyday practice. (Stolt 1994, 159, our translation)[1]

Trust in medicine as taught by the universities decreased. Quackery had been an alternative for long in Sweden, and it is reported that in the 19th century it was equally natural to seek help from a "wise woman" as it is to visit the doctor today (Ling 2004, 21). The period is referred to by Swedish scholars ironically as 'the golden era of public distrust and humbug medicine' (Fåhræus 1950, 102).

Whatever the connection might have been medical science advanced rapidly during the second half of the 1800s: from Louis Pasteur's 1859 suggestion that microorganisms may cause many human and animal diseases, Joseph Lister's 1867 publication "On the Antiseptic Principle in the Practice of Surgery", showing that disinfection reduces post-operative infections, to Robert Koch's 1882-1883 isolations of the microorganism responsible for tuberculosis and cholera. In short, in Sweden the decades before the regulation was put in place involved several breakthroughs in medical science; it was a time when medical science, finally, advanced to a position from which it could actually prove useful to medical practice. It is indeed interesting that those who explain the

[1] Approximate translation of the Swedish original: "[Landsortspraktikerna] prövade sig fram, och ännu runt 1850 hade de förvånansvärt liten nytta av de medicinska teorierna i sin vardag."

emergence of EBM also refer to medical breakthroughs in the late 1800s: "The 100-year period between 1885 and 1985 brought amazing medical breakthroughs" (Howick 2011, 11).

It makes sense to ponder what happened in Finland during this period. For centuries – until 1809 – Finland and Sweden were joined. Considerable overlap with regard to the requirement of science and proven experience between the two countries would come as no surprise. The Medical Society of Finland[2] (*Finska läkaresällskapet*) was founded in 1835 with the dual purpose of developing medical science and health care. It was followed by The Finnish Medical Society (*Duodecim*) in 1881, which aimed to develop medical science and practice in Finland.[3] Nowadays the overwhelming majority of Finnish physicians who are members of The Finnish Medical Association (*Lääkäriliitto*), founded in 1910, are committed to treating patients in accordance with science and proven experience through the code of medical ethics approved by the association's delegate committee (Lääkäriliitto

[2] We are unsure whether there is an official English translation, but "The Medical Society of Finland" is sometimes used when Finska läkaresällskapet is referred to in English contexts.

[3] The Medical Society of Finland was set up specifically for the Swedish speaking community of practice, whereas the explicit aim of Duodecim was to promote medical practice and uptake of medical science in Finnish. A third society, Suomen Lääkäriliitto (The Finnish Medical Association), was established in 1910. Many thanks to Matti Sintonen for generously helping us to navigate the various medical associations of Finland.

2014). Moreover, it is a legal requirement in Finland since 1994 that all health care personnel should only apply methods that there is proven experience of (since 2000 the same requirement holds for those in veterinary care).[4] Much of the development leading to the current role of VBE in Finland has happened after 1809. Thus the overlap is not a mere historical artefact. So what we refer to as the Swedish concept has a perfect match in Finland, although it is much more widespread in its Swedish applications. The motivation

[4] *Lag om yrkesutbildade personer inom hälso- och sjukvården* (1994/559) §15: ”En yrkesutbildad person inom hälso- och sjukvården skall i sin yrkesutövning tillämpa allmänt godtagna och beprövade metoder enligt sin utbildning, som han hela tiden skall försöka komplettera. I samband med yrkesutövningen skall en yrkesutbildad person inom hälso- och sjukvården opartiskt beakta den nytta och de eventuella olägenheter den medför för patienten.” See also *Lag om utövning av veterinäryrket* (2000/29) §13. Läkarens etiska regler, accepted by Suomen Lääkäriliitto (The Finnish Medical Association) in 1988 states in §V that: ”Läkaren skall upprätthålla och förkovra sina kunskaper och sitt kunnande och han skall endast rekommendera undersökningar och behandlingar som i enlighet med medicinsk vetenskap och beprövad erfarenhet anses effektiva och ändamålsenliga” (Saarni 2006, 11). A closely similar formulation occurs in the latest version of the codex, approved 2014 (see Lääkäriliitto 2014). The English translation of the relevant passage reads: ”A physician shall maintain and improve his knowledge and skills. He shall use and recommend only such examinations and therapies which medical knowledge and experience have shown to be effective and purposeful” https://www.laakariliitto.fi/en/ethics/.

behind the requirement in Finland is arguably the same as in Sweden since the two countries had so much in common during the concept's pre-history.[5]

In fact, moving outside of the strict domain of VBE-M one could claim that VBE is a Nordic concept. For instance, since 1998 psychologists in the Nordic countries are committed to VBE through *Yrkesetiska principer för psykologer i Norden*. These principles state: "Psykologen arbetar i enlighet med vetenskap och beprövad erfarenhet och eftersträvar en kontinuerlig professionell utveckling genom att inhämta mer och ny kunskap om den vetenskapliga och yrkesmässiga utvecklingen."
(Sveriges_Psykologförbund 1998, 6)[6]

Certainly, it was not only in the Nordic countries that it was acknowledged, around 1890, that medical science ought to guide medical decision-making. Failure to consider medical science in the medical profession was criticized in *The Boston Medical and Surgical Journal* too:

> … medical art without science is not only unprogressive, but almost inevitably becomes quackery. As soon as we treat our patients by rule of thumb, by tradition,

[5] It hasn't been possible to track the pre-history with sufficient certainty and accuracy within this project – at least not so far – but there appears to be similar ideas in 1800 century writings by, for instance, the father of pediatrics in Sweden, Nils Rosén; moreover, a predecessor from 1733 resembles in some respects the 1829 oath we have referred to above (Eklöf 2000).

[6]

> by dogmas, or by metaphysical axioms, we do injury to ourselves as well as to them. (Pye-Smith 1900, 173)

Still, if we (simplistically) compare occurrences of the expression "science" *("vetenskap"*) with occurrences of the expression "experience" (*"erfarenhet"*) in sources such as The Transactions of The Medical Society of Finland[7] (*Finska läkaresällskapets handlingar*) in the latter half of the 1800s we immediately find that, whereas quite a few reports contain the word "experience", there is less mention of "science". Hence it is understandable that further measures, such as the requirement of VBE-M, were put in place to ensure that doctors made more use of science in practice.

There is arguably a similar story to be told about the emergence of EBM exactly 100 years later. Here both Guyatt and Sackett report on the need to be sceptical vis-à-vis received medical wisdom (Howick 2011, chapter 2).

(3) However, it seems that experience of a specific kind – *proven* experience – was also identified as important as a result of the arrival of VBE-M. Somewhat paradoxically it seems that the development of relevant scientific evidence was accompanied by a corresponding development (or upgrading, or rating-up) in the role of evidence of a certain kind from experience as well. This is the third relevant comparison point between EBM

[7] There is, as far as we know, no official translation of *Finska läkaresällskapets handlingar* into English.

and VBE-M. Looking at the recent introduction of science and proven experience in the Swedish Educational Act (2010:800), we can see that Swedish schools and education authorities have developed a growing interest in proven experience. In particular, these discussions highlight the evidential relevance of experience within the professional collective. A third comparison between EBM and VBE-M can therefore be based on the way they deal with the relationship between two different types of evidence: evidence that is 'scientific' and evidence that is 'experienced based'.

Whether the prominence of proven experience is to be counted as a similarity between EBM and VBE-M depends on whether EBM and VBE-M have the same effect of rating up experience of the proven kind. We are ready to argue that they do, but this assessment depends on an assumption few advocates of EBM endorse. The assumption is that there is no strong link between EBM and science – or rather that the link, such as it is, is no stronger than that connecting EBM and proven experience.[8] In other words, the scientific classification here is not straightforwardly guaranteed by the use of certain methods or methodologies recommended by EBM such as, for example, randomised controlled trials. Much of what is regarded as being at the core of EBM could then equally well be classified as proven experience. To the extent that this assumption is accepted there is reason to think that EBM would simply prioritize experience of a certain kind. Advocates of EBM might

[8] For a related observation, see (Stoyanov, Machamer, and Shaffner 2012, 150): "In fact, they do nothing else but retell us the fragmented individual narrative, but presented in an ostensibly scientifically structured manner."

be dissatisfied with that implication because they wish to preserve the link between basic science and clinical research:

> Evidence-based medicine focuses on these systematic studies simply because they represent the most advanced stages of testing to ascertain whether the innovations of basic science work, how well they work, and for whom they work when applied in the clinical setting. Thus, evidence-based medicine is not in competition with basic science; rather it depends on it and builds on it. (Haynes et al. 1996, 196-97)

A clear dissimilarity – no matter how EBM is construed with respect to science and proven experience – would be that EBM *downgrades* certain kinds of science, such as science that is not based on RCTs (e.g. cohort studies), but this is not the case with VBE-M, at least not explicitly. In the next section we will look more closely at the way EBM and VBE-M handle the notion of evidence.

A further dissimilarity between EBM and VBE-M can be detected. The meaning of VBE-M varies with context among medical practitioners. (Persson and Wahlberg 2015) reports that the BE (or proven experience) component is sometimes taken to report a property of doctors and sometimes used to refer to a fact about how seriously a therapy has been tested in practice. By contrast, EBM seems fixed. Indeed if it were not fixed it would be difficult to understand the need for instruments such as the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework. GRADE is a framework

for synthesizing and rating the quality of evidence, and for providing clinical practice guidelines addressing alternative management options. It is used by many important actors in health care all over the world, including WHO and the Cochrane Collaboration.

To the extent that EBM introduces a new paradigm – see, for instance, Evidence Based Medicine Working Group (1992) – it ought to follow that something of fundamental importance remains fixed through various applications of EBM. Our hypothesis is that the primary candidate for this static role is the EBM position on evidence (see next section, and (Howick 2011, 4)).

3. EBM and VBE-M: The question of evidence

There is a clear sense in which medicine must always be based on some kind of evidence. If evidence is merely a ground for belief, there will not be anything new about EBM. The "evidence" in EBM, and in the "science/vetenskap" (V) and "proven experience/beprövad erfarenhet" (BE) of VBE-M, is all about what medical practitioners, or policy makers, can *justifiably* base their decisions on.

EBM is a procedure, or approach, that ensures, or perhaps maximises, justifiable decisions. VBE-M, as it stands, is a criterion for evaluating whether a decision is warrantable. This does not entail that VBE is only put to use post hoc. VBE can be used prospectively, too, as can be seen from the discussions in the *Journal of the Swedish Medical Association*

cited in (Persson and Wahlberg 2015) and in the oath taken by Finnish physicians – which requires that one to tries to advance proven experience in one's field.

So, what is this evidence that makes EBM different from medicine practised before 1990? There are two suggestions, both present in the prehistory of EBM. Archie Cochrane wrote in 1972:

> It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials. (quoted from Sønbø Kristiansen and Mooney 2004, 2)

The first is that EBM builds on the principle that all relevant evidence should be taken into account.

The second is that EBM builds on an idea of levels of evidence which not only identifies but also ranks the relevant kinds of evidence. For Cochrane in 1972 it was randomised controlled trials that constituted the relevant level. In the case of EBM, the Oxford Centre for Evidence-Based Medicine (CEBM) presents a comprehensive list of levels of evidence for different clinical questions. For therapy/prevention these are, from the top down: systematic review of RCTs, individual RCTs (all or none), systematic reviews of cohort studies, individual cohort studies, "outcomes" research, ecological studies, individual case-

control studies, case series, and last, expert opinions either without explicit critical appraisal or based on physiology, bench research or "first principles" (CEBM 2009).

It is obvious that evidence from basic medical science (physiological processes) and institutional or individual experience are not held in high regard.[9] The evidence on which decisions should be based is that deriving from clinical research. This evidence has high predictive value. An early presentation of EBM from the Evidence Based Medicine Working Group (1992), states that:

> Evidence-based medicine de-emphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision making and stresses the examination of evidence from clinical research. (EBMWG 1992, 2420)

Given this understanding of evidence, which is clearly based on clinical *research*, the question becomes what to do about clinical *experience*. Needless to say, clinical experience is often decisive for predictive purposes (see, for instance, Cartwright and Hardie 2012).

[9] Holly Andersen provides an interesting argument explaining why this is the case (Andersen 2012). In general the fact that our bodies are complex evolved systems makes it likely that relevant variables will be masked.

It is certainly plausible to say that EBM cannot be based *solely* on evidence from clinical research. Results from clinical research are not always there to be had, and where they are unavailable unsystematized clinical experience can be used as evidence:

> ... systematic attempts to record observations in a reproducible and unbiased fashion markedly increase the confidence one can have in knowledge about patient prognosis, the value of diagnostic tests, and the efficacy of treatment. In the absence of systematic observation one must be cautious in the interpretation of information derived from clinical experience and intuition, for it may at times be misleading. (EBMWG 1992, 2421)

This is not the situation with regard to VBE-M. Good therapeutic decision-making rests, according to VBE-M, on two *evidential* sources – science and proven experience. In EBM, by contrast, acknowledgement of the importance of clinical experience as evidence seems to be limited to cases where there are no relevant research findings.

This way of conceiving of EBM might not be shared by Swedish doctors. Trained as they are in thinking about VBE-M, it is natural for Swedish practitioners to assume that EBM has a place for BE as evidence, too (EBM replaces, as it were, the older idea of science (V), in VBE):

> It is a misunderstanding to assume that EBM no longer involves what we have called 'proven experience' … The right way to use personal experience is to contrast experience against the literature when a current problem is analysed. (Werkö et al. 2002, 3478-79, our translation)[10]

(Indeed, Professor Lars Werkö was one of the leading Swedish physicians. He became president of the Swedish Medical Society and ended his career as director of SBU (Hont 2009))

However, as we shall argue in the next section, the combination of EBM and BE is sometimes problematic. EBM's take on evidence, in what appears to be the most common version of EBM internationally (the *A-views* subsection, see below), is too restrictive to allow for full incorporation of BE.

4. EBM and VBE-M: Integrating science and experience

In an influential statement of EBM by Sackett et al. (1996) it is clear from the subtitle of the paper that *individual clinical expertise* is also important in EBM:

> Evidence based medicine: what it is and what it isn't
> *It's about integrating individual clinical expertise and the best external evidence*

[10] Det är ett missförstånd att tro att EBM inte längre skulle röra det vi kallat 'beprövad erfarenhet'. … Den rätta användningen av den egna erfarenheten borde vara att i samband med analys av ett aktuellt problem ställa erfarenheten mot vad litteraturen visar.

However, in connection with EBM it is sometimes unclear whether this integration means (A) that EBM consists of several parts, with evidence from clinical research being one part and clinical expertise being another, or (B) that EBM is one part of the total decision situation, with such things as clinical experience and patient preferences being other components. (Eddy 2005) introduces a somewhat similar distinction between evidence-based medicine and evidence-based guidelines, arguing that the former concept is built around individuals (decision-makers as well as patients) but could usefully be widened so as to include the latter phenomenon (often including multi-disciplinary teams). Our point, however, is that the A-view is "internally" problematic.

4.1 A-views

(A) is clearly the more common version of EBM. A number of introductions to EBM present flow charts like that in Figure 1:

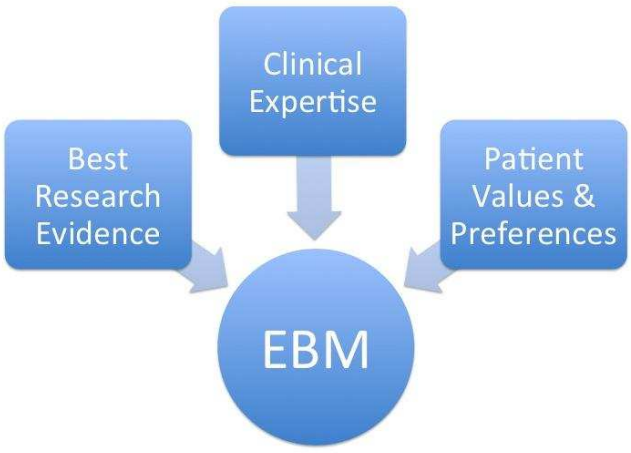


Figure 1: The A-view, adapted from

http://guides.mclibrary.duke.edu/c.php?g=158201&p=1036021 (downloaded 1-Feb 2016)

An influential position paper remarks that:

> Initially, evidence-based medicine focused mainly on determining the best research evidence relevant to a clinical problem or decision and applying that evidence to resolve the issue. This early formulation de-emphasised traditional determinants of clinical decisions, including physiological rationale and individual clinical experience. (Haynes, Devereaux, and Guyatt 2002)

These remarks concern the very first EBM-formulations. Sackett et al. (1996) is conceived as the original attempt to integrate evidence and clinical experience in a better way. Sackett et al. (1996) seemingly advocate a version of the A-view. The position paper continues:

> Evidence based medicine is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research. (Haynes, Devereaux, and Guyatt 2002)

According to this statement, EBM includes both evidence *and* clinical experience. EBM does not complement clinical experience; it includes it. However, it should be noted that Sackett et al. (1996) discuss evidence only in connection with external clinical experience, which is to say research. (Haynes, Devereaux, and Guyatt 2002) picture the position in Sackett et al. (1996) in a way that makes it very similar to the A-view shown in Figure 1 above.

It is said that the "concepts of evidence-based medicine are evolving as limitations of earlier models are addressed" (Haynes, Devereaux, and Guyatt 2002). However, in our view this has not led to the abandonment of the A-view. That view is perhaps even more clearly relied on in later formulations of EBM – e.g. in what is often presented as the "contemporary definition":

> … the integration of the best research evidence with clinical expertise and patient values. (Sackett et al. 2000, 71)

Now, as we have already touched upon, the A-view makes it difficult to talk about evidence as something other than *research* evidence. As (Haynes, Devereaux, and Guyatt 2002) puts it:

> Evidence-based medicine recognises that such evidence is not "created equal" and provides detailed guides for finding the most rigorous and pertinent evidence for a specific clinical decision.

As a consequence the use of clinical experience as evidence has been downplayed in later developments of EBM, and clinical experience is nowadays almost exclusively mentioned as that which is needed to implement scientific evidence in a specific decision context. An example of this is the entry on "Making a decision" on the CEBM website, www.cebm.net. Here, a decision is made by:

> Incorporating the findings of valid, important and applicable research with your patient values and preferences and your clinical expertise to arrive at the right decision about their individual health care. (CEBM 2016)

Duke University Medical Center, from whose work we adapted the flow chart in Figure 1, presents the issues in a similar way:

> The evidence, by itself, does not make the decision, but it can help support the patient care process. The full integration of these three components into clinical decisions enhances the opportunity for optimal clinical outcomes and quality of life. (http://guides.mclibrary.duke.edu/c.php?g=158201&p=1036021 (downloaded 1-Feb 2016))

In other words, the original idea behind EBM highlights the need to *integrate* research findings with individual clinical expertise, but on the A-view it is clear that this integration cannot be one in which two types of *evidence* are integrated, since that would violate EBM's paradigmatic view of evidence. Remaining within the paradigm might work in some cases, for certain types of clinical experience (proven experience of a certain kind), but normally the difference between evidence from the two sources would be too pronounced for anything but research evidence to count, according to EBM. This creates considerable tension within A-views, since they also wish to acknowledge the role of other kinds of "information":

> We title this component of clinical decisions 'research evidence' to distinguish it from other forms of information that have always been part of clinical decisions,

> such as the patient's history, physical findings, diagnostic tests, circumstances, and stated preferences. (Haynes, Devereaux, and Guyatt 2002)

4.2 B-views

The B-view, where EBM is one part of the total decision situation, also has advocates. It is perhaps not surprising that Swedish perspectives sometimes express B-views, since these are much easier to interpret in terms of VBE. For example, in a much quoted passage in a letter to a physician, the Swedish National Board of Health and Welfare explains:

> In the exercise of her profession, the medical doctor must take account of both science and proven experience. [...] When a new method is introduced, proven experience of it is trivially lacking, and the scientific evidence can suffice for acceptance […]. At other times, long clinical experience might be the dominating evidence in favour of accepting the medical treatment whereas theoretical and/or experimental evidence for its effectiveness might be lacking.

(Asplund 2001) presents a picture captured in the following flow chart:

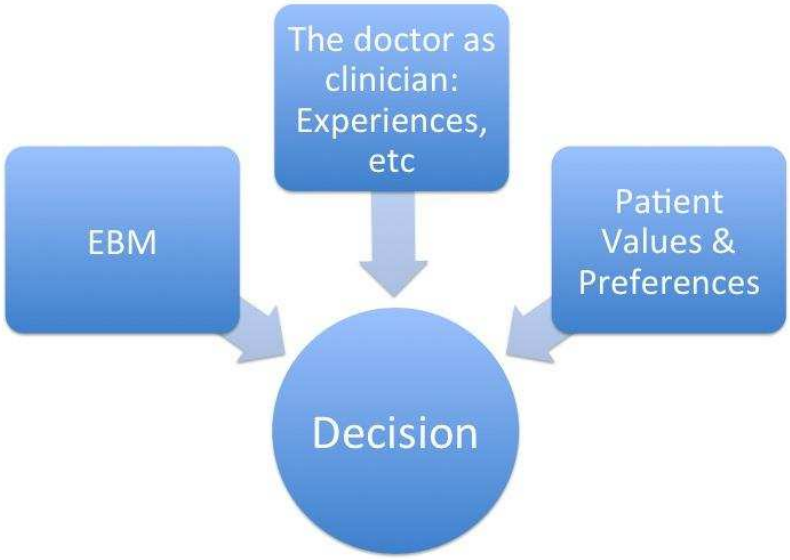


Figure 2: The B-view, modified from Asplund (2001)

The B-view is not a uniquely Swedish phenomenon. For instance, Haynes et al. 1996, p. 196, defines evidence-based medicine thus:

> Evidence-based medicine is the conscientious and judicious use of current best evidence from clinical care research in the management of individual patients.

At first blush this definition seems very similar to those we have referred to as A-views, and that is probably how it was intended to be presented. Here too, however, it might be argued that EBM is but one part of a larger decision context, the management of individual patients. But this is not normally how proponents of EBM picture it. They distinguish

between early and updated models (Haynes, Devereaux, and Guyatt 2002), and this would be one of the earliest – a model that "de-emphasised traditional determinants of clinical decisions, including physiological rationale and individual clinical experience." The alternative B-view reading would be that the early models present EBM as one component of the decision.

On a B-view it is much easier to understand EBM and its evidential levels. "The doctor as clinician" (see Figure 2) can bring evidence into the decision as well, but it is not the type of evidence EBM speaks of, which concerns research findings only. On the A-view the clinician's expertise is only a means of applying research evidence to a particular case. That expertise does not provide any additional evidence. On the B-view, by contrast, the individual and collective clinical experience that the clinician adds to the decision basis qualifies as relevant evidence too. Consequently, according to the B-view EBM does not really set a standard for decision-making (see above). Its capacity to help doctors make better decisions is clearly weakened. EBM becomes much more of a *partial* tool for decision-making than advocates of EBM normally assume.

4. Concluding remarks

Advocates of EBM struggle to model the way evidence-based medicine should "[integrate] individual clinical expertise and the best external evidence" (Sackett et al. 1996). We have argued that the formulations and models available in the literature today are variations on a common theme. On these A-views EBM describes all that goes on in a specific *justifiable*

medical decision. A-views inevitably create tensions in the concept of evidence they require.

For that reason alone B-views are of interest. On a B-view EBM describes just one component of the decision situation (a component usually based on evidence from clinical research). Together with other types of evidence, EBM leads to a justifiable clincial decision, but it does not describe the decision itself. The B-view is inspired by a 100-years older version of EBM, a Swedish standard that requires medical decision-making and practice to be consistent with 'science and proven experience'.

In sum, the Swedish concept of 'science and proven experience' clearly resonates with several characteristics of evidence-based medicine. Like EBM it focuses on evidence (rather than opinion), on science, and on the need for integration.

However, the Swedish concept also differs from the concept of evidence-based medicine in that it clearly identifies two sources of evidence as special: science (vetenskap) and proven experience (beprövad erfarenhet). Comparing EBM and VBE, one is struck by the relative clarity of the Swedish notion.

[Acknowledgements to be inserted after review]

References

"Patient Safety Act (2010:659)."
2010:800, Skollagen. Stockholm: Norstedts juridik.
Andersen, Holly. 2012. "Mechanisms: what are they evidence for in evidence-based medicine?" *Journal of Evaluation in Clinical Practice* 18:992-999.
Asplund, Kjell. 2001. "Den evidensbaserade medicinen är nödvändig men inte tillräcklig: Bör kompletteras inom områden där det vetenskapliga underlaget är svagt." *Läkartidningen* 98 (37):3898-3901.
Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-based policy – a practical guide to doing it better*. New York: Oxford University Press.
CEBM. 2009. "Oxford Centre for Evidence-based Medicine – Levels of Evidence (March 2009)." Accessed 2016 February 25. http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/.
CEBM. 2016. "Making a Decision." Accessed 2016 February 25. http://www.cebm.net/category/ebm-resources/tools/making-a-decision/.
Danek, K. 1969. "Towards a history of health." In *Nordisk medicinhistorisk årsbok*.
EBMWG. 1992. "Evidence-based medicine. A new approach to teaching the practice of medicine." *JAMA* 268 (17):2420-2425.
Eddy, David M. 2005. "Evidence-based medicine: A unified approach." *Health Affairs* 24 (1):9-17.
Eklöf, Motzi. 2000. *Läkarens Ethos: Studier i den svenska läkarkårens identiteter, intressen och ideal 1890-1960*. Vol. 216, *Linköping Studies in Arts and Science*. Linköping.
Fåhræus, Robin. 1950. *Läkekonstens historia, del III*. Stockholm: Bonniers.
Haynes, R. Brian, P. J. Devereaux, and Gordon H. Guyatt. 2002. "Clinical expertise in the era of evidence-based medicine and patient choice." *Evidince-Based Medicine* 7 (2):36-38.
Haynes, R. Brian, David L. Sackett, J.A. Muir Gray, D.J. Cook, and Gordon Guyatt. 1996. "Transferring evidence from research into practice: 1. The role of clinical care research evidence in clinical decisions." *Evidence-Based Medicine* 1 (7):196-198.
Hont, Gabor. 2009. "Svensk sjukvårds nestor Lars Werkö död." *Läkartidningen* 106 (42):2672.
Howick, Jeremy. 2011. *The Philosophy of Evidence-Based Medicine*, *BMJ Books*: Wiley-Blackwell.
Ling, Sofia. 2004. *Kärringmedicin och vetenskap*. Edited by Torkel Jansson, Jan Lindegren and Maria Ågren. Vol. 212, *Studia Historica Upsaliensa*. Uppsala: Acta Universitatis Upsaliensis.
Lääkäriliitto. 2014. *Läkarens etiska regler*. Finland: Lääkäriliitto.
Persson, Johannes, and Lena Wahlberg. 2015. "Vår erfarenhet av beprövad erfarenhet: några begreppsprofiler och ett verktyg för precisering." *Läkartidningen* 112 (49).
Porter, Roy. 1995. *Disease, Medicine and Society in England 1550-1860*. 2nd edition ed: Cambridge University Press.
Pye-Smith, Philip H. 1900. "Medicine as a Science and Medicine as an Art." *The Boston Medical and Surgical Journal* 143 (8):173-177. doi: doi:10.1056/NEJM190008231430801.
Rosenberg, William, and Anna Donald. 1995. "Evidence based medicine: an approach to clinical problem-solving." *BMJ* 310:1122-1126.

Saarni, Samuli, ed. 2006. *Medicinsk etik*. 6 upplagan ed. Helsingfors: Finlands Läkarförbund.

Sackett, David L., William M.C. Rosenberg, J.A. Muir Gray, R. Brian Haynes, and W. Scott Richardson. 1996. "Evidence based medicine: what it is and what it isn't." *BMJ* 312:71-72.

Sackett, David L., Sharon E. Straus, W. Scott Richardson, William Rosenberg, and R. Brian Haynes. 2000. *Evidence-Based Medicine: How to Practice and Teach EBM*. Edinburgh: Churchill Livingstone.

Stolt, Carl-Magnus. 1994. *Den beprövade erfarenheten: Medicinsk idéhistoria och läkekonst i Boråsbygden 1780-1900*. Borås: Norma.

Stoyanov, Drozdstoj St., Peter K. Machamer, and Kenneth F. Shaffner. 2012. "Rendering clinical psychology an evidence-based scientific discipline: a case study." *Journal of Evaluation in Clinical Practice* 18:149-154.

Sveriges_Psykologförbund. 1998. Yrkesetiska principer för psykologer i Norden. edited by Sveriges Psykologförbunds kongress 1998.

Sønbø Kristiansen, Ivar, and Gavin Mooney. 2004. *Evidence-Based Medicine – in its place*: Routledge.

Werkö, Lars, Kjell Asplund, Peter Aspelin, Mona Britton, Mats Eliasson, Jean-Luc af Geijerstam, and Sten Thelander. 2002. "Två år med EBM i Läkartidningen: Klinisk forskning och rutinsjukvård har närmat sig varandra." *Läkartidningen* 99 (36):3478-3482.

No Free Lunch Theorem, Inductive Skepticism, and the Optimality of Meta-Induction

Word count: 4986

Gerhard Schurz

Abstract: The no free lunch theorem (Wolpert 1996) is a radicalized version of Hume's induction skepticism. It asserts that relative to a uniform probability distribution over all possible worlds, all computable prediction algorithms – whether 'clever' inductive or 'stupid' guessing methods (etc.) – have the same expected predictive success. This theorem seems to be in conflict with results about meta-induction (Schurz 2008). According to these results, certain meta-inductive prediction strategies may dominate other (non-meta-inductive) methods in their predictive success (in the long run). In this paper this conflict is analyzed and dissolved, by means of probabilistic analysis and computer simulation.

1. The Optimality of Meta-Induction: A Solution to the Problem of Induction?

In Schurz (2008) a new account to the problem of induction has been developed that is based on the optimality of meta-induction. The account agrees with Hume's skeptical insight that it is impossible to demonstrate a priori that induction is *reliable* in the sense that it is predictively more successful than random guessing. Such a demonstration is impossible without assuming that the actual world possesses a certain amount of regularity. Reichenbach (1949, §91) argued that it is at least possible to demonstrate a priori that induction

is *optimal*, i.e., is the best what we can do for the purpose of predictive success. Results in formal learning show, however, that it is not possible to demonstrate optimality at the level of *object-induction*, that is, of induction applied to the task of predicting events in arbitrary possible worlds (cf. Skyrms 1975, ch. III.4). In contrast, what the account of *meta-induction* attempts to show is that induction is optimal if it is applied at the meta-level of competing prediction methods. The meta-inductive strategy tracks the success rate of all prediction methods whose predictions are *accessible* and predicts an optimal weighted average of the predictions of those methods that were most successful so far. Based on results in mathematical learning theory (Cesa-Bianchi 2006), Schurz (2008) proved that there exists a particular weighting method, called *attractivity-weighting*, which grants the meta-inductivist a predictive success rate that is in the long run at least as high as that of every other prediction method that is accessible to the meta-inductivist, even if their success rates are permanently changing in an irregular way. Since the restriction to accessible methods is crucial for the optimality theorem, Schurz and Thorn (2016) call this kind of optimality *access-optimality*. Remarkably, the access-optimality of meta-induction holds in *all* possible worlds, even in 'chaotic' ones in which event frequencies do not converge against limits or in 'paranormal' worlds which host clairvoyants.

Technically the account of meta-induction is based on the notion of a prediction game:

Definition 1. A *prediction game* is a pair ((e),Π) consisting of:

(1.) An infinite sequence (e) := (e_1,e_2,…) of events e_n, coded by real numbers between 0 and 1, possibly rounded according to a finite accuracy. For example, (e) may be a sequence

of daily weather conditions, football game results, or stock values. In what follows $Val \subseteq [0,1]$ denotes the value space of possible events $e_n \in Val$. Each time n corresponds to one round of the game.

(2.) A finite set of prediction methods or 'players' $\Pi = \{P_1,\ldots,P_m,MI\}$ (in what follows we identify 'methods' with 'players'). In each round it is the task of each player to predict the next event of the event sequence. "MI" signifies the meta-inductivist and the other players are the 'non-MI-players' or 'candidate methods'. They may be real-life experts, virtual players implemented by computational algorithms, or even 'clairvoyants' who can see the future in 'para-normal' possible worlds. It is assumed that the predictions of the non-MI players are accessible to the meta-inductivist. Moreover, they are elements of $Val \subseteq [0,1]$.

The *predictive success rate* of a method P is defined by means of the following chain of definitions:

– $pred_n(P)$ is the prediction of *player* P *for* time n delivered *at* time n−1,

– the deviation of the prediction $pred_n$ from the event e_n is measured by a normalized loss function, $loss(pred_n,e_n) \in [0,1]$,

– $score(pred_n,e_n) =_{def} 1-loss(pred_n,e_n)$ is the *score* obtained by prediction $pred_n$ of event e_n,

– $abs_n(P) =_{def} \Sigma_{1\leq i\leq n}\, score(pred_i(P),e_i)$ is the *absolute* success achieved by player P until time n, and

– $suc_n(P) =_{def} a_n(P)/n$ is the *success rate* of player P at time n.

The *natural* loss-function is defined as $|pred_n-e_n|$. The optimality theorem holds below for

all *convex* loss functions, which means that the loss of a weighted average of two predictions is not greater than the weighted average of the losses of two predictions. In what follows we assume convex loss functions; they comprise a large variety of loss functions including all linear, polynomial, or exponential functions of the natural loss function.

'Possible worlds' are identified with prediction games. A special case are *binary* games whose events and predictions are elements of {0,1}. For binary games the natural loss function coincides with the zero-one loss: $\text{loss}_{1\text{-}0}(\text{pred},e) = 0$ if pred = e, and otherwise = 1.

The simplest meta-inductive strategy is called "Imitate-the-best" and predicts what the presently best non-MI player predicts. It is easy to see that this meta-inductive method cannot be universally access optimal: Its success rate breaks down when it plays against non-MI methods that are *deceivers*, which means that they lower their success rate as soon as their predictions are imitated by the meta-inductivist (cf. Schurz 2008, sec. 4). A realistic example is the prediction of stock values in a 'bubble economy': Here the prediction that a given stock will yield a high rate of return leads many investors to put their money on this stock and by doing so they cause it to crash. Nevertheless there exists a meta-inductive strategy that is provably universally optimal. This strategy is called *attractivity-weighted meta-induction*, abbreviated as wMI, and is defined as follows:

Definition 2. The predictions of wMI (attractivity-weighted meta-induction) are defined as

$$\text{pred}_{n+1}(\text{wMI}) =_{\text{def}} \frac{\sum_{1\le i\le m} \text{at}_n(P_i)\cdot \text{pred}_{n+1}(P_i)}{\sum_{1\le i\le m} \text{at}_n(P_i)}\text{, where}$$

– $\text{at}_n(P_i)$ is the attractivity of a player P_i for wMI at a given time n, defined as

$at_n(P_i) =_{def} suc_n(P_i) – suc_n(wMI)$, if this expression is positive; else $at_n(P_i)=0$, and

– if n=1 or the denominator is zero, wMI's prediction is a random guess.

Let "$maxsuc_n$" denote the non-MI-players' maximal success rate at time n. Then the optimality theorem for wMI (proved in Schurz 2008, sec. 7, theorem 4) asserts:

Theorem 1. (Universal access-optimality for wMI):

For every prediction game ((e), $\{P_1,\ldots,P_m,wMI\}$) the following holds:

(1.1) (Short run:) $(\forall n\geq 1:)\ suc_n(wMI) \geq maxsuc_n - \sqrt{m/n}$.

(1.2) (Long-run:) $suc_n(wMI)$ (strictly) converges to the non-MI-players' maximal success for $n\rightarrow\infty$.

According to theorem (1.2) attractivity-weighted meta-induction is long-run optimal for *all* possible event sequences and sets of accessible prediction methods. The only proviso is that the set of accessible methods is finite, which is a realistic assumption for cognitively finite beings. In the short run, weighted meta-induction may suffer from a possible loss, compared to the leading player. This loss (which is also called wMI's 'regret') is caused by the fact that wMI must base her prediction of the next event on the *past* success rates of the candidate methods, and the hitherto most attractive methods may perform badly in the prediction of the *next* event. Fortunately theorem (1.1) states a worst-case upper bound for this loss, which is small if the number of competing methods, m, is small compared to the number of rounds, n, and which converges quickly to zero when n grows large.

Theorem 1 applies to prediction games with real-valued as well as binary (or discrete) events. Even if the events are binary wMI's predictions are real-valued (because proper weighted averages of 0s and 1s are real-valued). How can the optimality result of theorem 1 be transferred to binary games whose predictions must be binary? There are two methods by which this can be done:

(1.) Randomization, rwMI (cf. Cesa-Bianchi and Lugosi 2006, sec. 4.1): Here one assumes that rwMI predicts e_n=1 with a probability (P) that equals the optimal real-valued prediction of wMI, i.e., P($pred_n$(rwMI) =1) = $pred_n$(wMI). This method is not entirely general since it presupposes that the events are probabilistically independent from rwMI's choice of prediction.

(2.) Collective meta-induction, cwMI (Schurz 2008, sec. 8): Here a *collective* of meta-inductivists approximates real-valued predictions by the mean value of their binary predictions. Their mean predictive success rate approximates provably the success rate of the optimal method wMI. Assuming that the cwMIs are *cooperators* and share their success, every individual member of the collective is predictively optimal.

Theorem 1 establishes the following a priori justification of attractivity-weighted meta-induction: In all environments it is reasonable – in *addition* to searching for good object-level methods – to apply the strategy wMI, as this can only improve but not worsen one's success in the long run. Note that by itself this justification does not entail anything about the rationality of object-level induction: it may well be that we live in a world in which a method different from object-induction is predictively superior. However, it seems that the a priori justification of meta-induction give us the following a posteriori justification of

object-induction: to the extent that (a particular version of) object-induction was so far the most successful prediction strategy, it is meta-inductively reasonable to continue favoring (this particular version of) object-induction.

Theorem 1 asserts the optimality but not the dominance (in the long run) of attractivity-based meta-induction. Thus there may exists other methods, different from wMI, that are likewise long-run optimal. In fact one can prove that there are certain variants of wMI that are long-run optimal and have short-run advantages in certain and disadvantages in other environments. So wMI is cannot be universally long-run dominant. Nevertheless, the following restricted dominance result for wMI follows from theorem 1:

Theorem 2. (Dominance for wMI):

(2.1) wMI dominates every prediction method that is not universally long-run optimal. In other words, for every such method M there is a prediction game containing wMI and M in which wMI's long-run success rate exceeds that of M.

(2.2) Not universally long-run optimal are, for example, all *independent* non-clairvoyant methods, that is, methods that can learn only from observations of past events, but not from the predictions of other methods.

Proof of theorem 2: Theorem (2.1) is an immediate consequence of theorem 1 and the definition of "optimality". The proof of theorem (2.2) goes as follows: Let M be an independent method based on a function f that maps each n-tuple of past events $(e_1,\ldots,e_n) \in Val^n$

into a prediction $pred_{n+1} \in Val$. We define an M-adversarial event sequence (e') as follows: $e'_1 = 0.5$, and $e'_{n+1} = 1$ if $f(e'_1,\ldots,e'_n) \leq 0.5$; else $e'_{n+1} = 0$. Moreover we identify the predictions of the perfect (e')-forecaster M' with the so-defined sequence, i.e., $pred_n(M') = e'_n$ (note that if f is computable, M' is so, too). In the prediction game ((e'),{M,M',wMI}) the success rate of M can never exceed 1/2, that of M' is always 1 and that of wMI converges to 1 (by theorem 1). This proves theorem 2. Q.E.D.

Theorem 2 is crucial for the next sections in which we confront the optimality of meta-induction with the no free lunch theorem.

2. Radical Inductive Skepticism: The No Free Lunch Theorem

Wolpert's (in)famous no free lunch theorem (Wolpert 1996) is a radicalized version of Hume's inductive skepticism for theoretical computer science. The theorem applies to prediction methods that can be represented as computable functions from past observations to predictions, so called *learning algorithms* (thus, clairvoyance is excluded). The theorem is often expressed by the assertion that for each pair of prediction methods, the number – or in the infinite case the probability – of possible worlds (event sequences) in which the first method outperforms the second is precisely equal to the number (or probability) of worlds in which the second method outperforms the first. We call this assertion the *strong* version of Wolpert's theorem, because it presupposes a 'homogeneous' loss function:

Theorem 3. Strong no free lunch theorem (Wolpert 1996, 1354f, theorems 1, 3):
For every possible loss value c, the probability of worlds in a which prediction method leads to a loss of c is the same for all possible prediction methods, *provided* one assumes
– (a) a *state-uniform* prior probability distribution, that is, a uniform distribution over all possible event sequences (or *states* of the world), and
– (b) a homogeneous loss function, in the sense that for all possible loss values c, the number of possible events e ∈Val for which a prediction pred∈Val leads to a loss of c is the same for all possible predictions pred∈Val.

The requirement of a homogeneous loss function very strong: It is only satisfied if events *and* predictions are binary, or more generally, if they are discrete with a zero-one loss function. Under this assumption homogeneity is obvious: If the value space has k elements, then for every pred∈Val the number of possible events e∈Val that lead to a loss of 1 is obviously k−1, and the number of events that lead to a loss of 0 is one. In contrast, in prediction games with real-valued predictions the homogeneity requirement fails. In the binary case, for example, the number of events which lead to a loss of 1 is one for the two predictions pred =1 and pred = 0, but zero for the prediction pred = 0.5.

Homogeneous loss functions are a clear restriction of the strong no free lunch theorem, since, as we have seen, real-valued predictions can be implemented even in binary games, either by randomized binary predictions or by a cooperative collective of binary forecasters. There is, however, a weak version of the no free lunch theorem (mentioned by Wolpert

1996 on p. 1354) which applies to real-valued predictions over binary or discrete events and assumes what we call a "weakly homogeneous" loss function:

Theorem 4. Weak no free lunch theorem (Wolpert 1996, 1354):
The probabilistically expected success of every possible prediction method is equal to the expected success of random guessing or of every other prediction method, provided one assumes

– (a) a *state-uniform* prior probability distribution, and

– (b) a weakly homogeneous loss function, in the sense that for every possible prediction pred ∈ Val the *sum* of pred's losses over all possible events e∈Val is the same ($\forall$pred∈Val: $\Sigma_{e\in Val}$loss(pred,e) = a constant c*).

For binary events with real-valued predictions and a natural loss function weak homogeneity is satisfied, since for every prediction pred∈[0,1], loss(pred,1) + loss(pred,0) = 1−pred + pred = 1.

For prediction games with real-valued events, most loss functions (including all convex ones) are not even weakly homogeneous. Here "free lunches" are possible in the sense that not all prediction methods have the same expected success, relative to a state-uniform probability distribution.

In this paper we focus on prediction games with discrete events and real-valued predictions, to which the weak no free lunch theorem applies. The framework in which Wolpert proves his theorems are not prediction games, but learning algorithms that map training

sets into predictions of test items. But since a prediction game can be considered as an iterated procedure of selecting a training set of n events and predicting the event at test item n+1, Wolpert's result applies straightforwardly to prediction games.

Theorem 4 asserts that *every* possible prediction method – be it an intelligent inductive one, a crazy anti-inductive one, or a stupid one that always predicts the same value – has the same expected predictive success relative to a state-uniform prior distribution. For all induction-friendly philosophical programs, including the program of meta-induction, this result seems to be devastating. How is it possible? In what follows we give a brief explanation of Wolpert's theorem in terms that are philosophically more familiar than his own "extended Bayesian framework".

Wolpert's theorem is a far-reaching generalization of a straightforward result about the prediction of binary sequences. For this application the strong no free lunch theorem amounts to the following: However a prediction function f, with $pred_{n+1} = f((e_1,\ldots,e_n)) \in \{0,1\}$, is defined, there are as many sequences of a given length k>n extending $(e_1,\ldots,e_n)$ that verify f's prediction $pred_{n+1}$ as there are sequences that falsify it. Thus by attaching an equal probability to every possible sequence the expected score of each prediction function will be 1/2. More generally speaking, this result is an immediate consequence of an (in)famous result in probability theory which can be found (among other authors) in Carnap (1950, 564-566) or Howson and Urbach (1996, 64-66). The result can be expressed as follows:

Theorem 5. (Carnap 1950, 564-566):

Let P be a state-uniform prior probability (density) distribution over (the Borel algebra over)[1] the set of all infinite binary sequences, $\{0,1\}^{\omega}$. Then P has the following two 'radically non-inductive' properties:

(a) P assigns the same conditional probability to each event $e_n \in \{0,1\}$ independently of the preceding events $(e_1,\ldots,e_{n-1})$ of the sequence. Thus, P is an IID (independent identical distribution) with $P(1) = P(0) = 1/2$.

(b) P assigns a probability of one to the class of sequences with a limiting frequency of 1/2 and a probability of zero to all other possible limiting frequencies; this follows from (a) by the strong law of large numbers.

3. No Free Lunch and Meta-induction – a Conflict?

We now turn to the relation between the weak no free lunch theorem and theorem 2 about meta-induction. The no free lunch theorem applies not only to object-level prediction methods, but also to all meta-strategies, given that they are applied to a *fixed* set of independent prediction methods – for the reason that every combination of a finite number of prediction algorithms is itself a prediction algorithm. So the puzzling question arises: If the

[1] P yields Carnap's confirmation function $c^{\dagger}$. Technically, $\{0,1\}^{\omega}$ is represented by the interval [0,1] of real numbers in binary representation (see fig. 1 below). P over the Borel algebra Bo([0,1]) is defined by the integrals of an assumed density function D_P over [0,1].

no free lunch theorem is true, how can it be that attractivity-weighted meta-induction, when applied to a fixed set of independent prediction algorithms, is dominant in comparison to certain other methods, as stated in theorem 2? Is this a contradiction?

Our answer to this question in regard to the long run perspective can be summarized as follows: No, the contradiction is only apparent. It is indeed true that there exist many wMI-accessible methods whose predictive success rate is (in the long run) strictly smaller than that of wMI in some worlds (event sequences)[2] and never greater than that of wMI in any world – let us call these methods M_{inf} (for "inferior"). Nevertheless the state-uniform expectation values of the success rates of wMI and M_{inf} are equal, because the state-uniform distribution that Wolpert assumes assigns a probability of zero to all worlds in which wMI dominates M_{inf}; so these worlds do not affect the probabilistic expectation value.

Let us elaborate on this connection. The major difference between the account of meta-induction and Wolpert's extended Bayesian account is this: While the former account is independent from any assumed prior distribution over possible event sequences, Wolpert's result depends on a particular prior distribution, the *state-uniform* distribution. Wolpert seems to assume that this distribution is epistemically privileged. Reasonable doubts can be raised here, because the state-uniform distribution is induction-hostile. A proponent of this distribution believes with probability 1 a priori that the binary event sequence she is

[2] Generally speaking possible worlds are identified prediction games. But in the given context we assume a fixed set of prediction methods, whence possible worlds can be identified with event sequences.

going to predict (a) has a limiting frequency of 1/2 and (b) is non-computable. Fact (a) follows from theorem 5, and (b) from the fact that there are uncountably many sequences, but only countably many computable ones. However, the event sequences for which an intelligent prediction method can be better than random guessing or any other stupid method are precisely those event sequences that *do not* fall into the intersection of classes (a) or (b). To make this point explicit: For random sequences with a limiting frequency of 1/2, all combinations of independent methods must have the same success rate as random guessing, i.e. 1/2. The only possibility for these sequences to be predictable is that they are computable by an internal regularity, but this possibility has probability zero, too.

In conclusion, proponents of a state-uniform prior distribution are strongly biased: they are a priori certain that the world is irregular so that induction cannot have any chance. We suppose that adherents of a more induction-friendly view, for example Bayesians in the ordinary (not Wolpertian) sense, will regard a state-uniform prior distribution as highly "unnatural". Instead of a state-uniform distribution they typically prefer a uniform distribution over all possible limiting frequencies; we call such a distribution a *frequency-uniform* distribution. It is well known that frequency-uniform distributions are highly induction-friendly: from them on can derive Laplace's rule of induction, $P(e_{n+1} = 1 \mid f_n(1) = \frac{k}{n}) = \frac{k+1}{n+2}$, where "$f_n(1)$" denotes the frequency of 1's among the first n events (cf. Carnap 1950, 568). In computer science, Laplace's rule has been generalized by Solomonoff (1964, sec. 4.1), who proved that if the prior probability of a sequence is inversely proportional to its *algorithmic complexity*, then Laplace's rule of induction is valid.

The precise relation between prior distributions over the space of possible infinite sequences and corresponding distributions over the space of possible limiting frequencies (or classes of sequences with the same frequency) is displayed in figures 1 and 2 below. As usual, infinite 0-1-sequences are represented as real numbers between 0 and 1 in binary representation (e.g., 0.0110…) and ordered according to their numerical size. In this way, the state-uniform distribution over possible sequences is represented as a uniform density over the interval [0,1]. Fig. 1 presents the transformation of this distribution into the corresponding distribution over possible limiting frequencies, with the result that a uniform distribution over [0,1] viewed as space of sequences is transformed into a maximally dogmatic distribution (an infinite density peak) over [0,1] viewed as space of frequency limits.

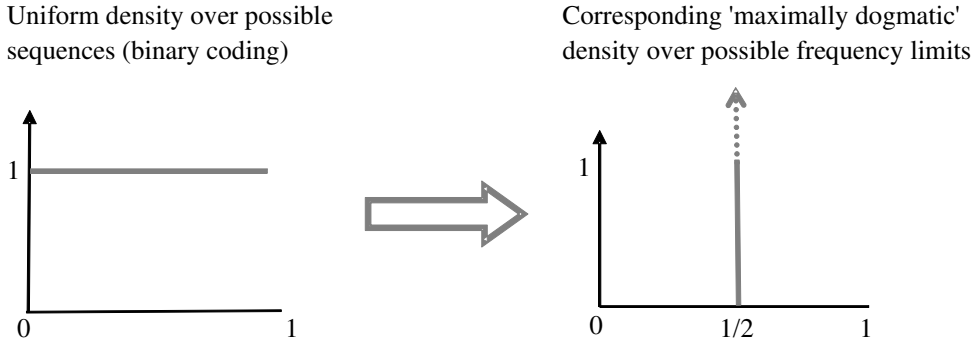


Figure 1. Transformation of a state-uniform into a frequency-uniform distribution.

Fig. 2 (below) illustrates the inverse transformation. The upper part of fig. 2 shows what happens to a frequency-uniform distribution over [0,1], if it is transformed into a distribution over [0,1] viewed as space of possible sequences. The resulting distribution becomes

non-continuous and entirely disrupted: in every finite interval $I \subseteq [0,1]$ it increases infinitely often to a positive value and falls back to zero.[3] It follows that a state-uniform prior distribution makes Bayesian converge results impossible, because all these results presuppose a (not necessarily uniform but) *continuous* prior distribution over the possible frequencies (cf. Earman 1992, 141ff). Thus "outwashing of priors" is impossible for state-uniform prior distributions. The lower part of fig. 2 displays Solomonoff's result (1964) which states that the frequency-uniform probability of a (finite or infinite) sequence decreases exponentially with its algorithmic complexity c(s): $P(s) \sim 2^{-c(s)}$. Thus sequences with lower complexity have a higher frequency-uniform probability than those with high complexity. In conclusion, a frequency-uniform distribution is strongly biased in regard to less complex (more regular) sequences.

So which prior distributions are more natural, state-uniform ones or frequency-uniform ones? In our eyes, this question has no reasonable answer because all prior distributions are subjective and biased in some respect. We regard it as a great advantage of the optimality of meta-induction that it holds regardless of any assumed prior probability distribution. For a frequency-uniform prior distribution the probability of worlds in which meta-induction dominates random guessing is close to one. For a state-uniform prior the probability of

---

[3] To see this, let r_1 and r_2 ($r_2 > r_1$) be two infinite sequences represented as binary real numbers r_1 = "0.00...(n times zero)11...(one forever)" and r_2 = "0.00...(n−1 times zero)11...(one forever)". Their complexity is minimal. The class of sequences lying between r_1 and r_2 contains sequences with all complexities between the minimal one and the maximal one, which is possessed by sequences with frequency limit 1/2. So the density climbs up and down between minimal and maximal complexity in the interval $[r_1,r_2]$. Since this holds for every arbitrary small interval, the claim follows.

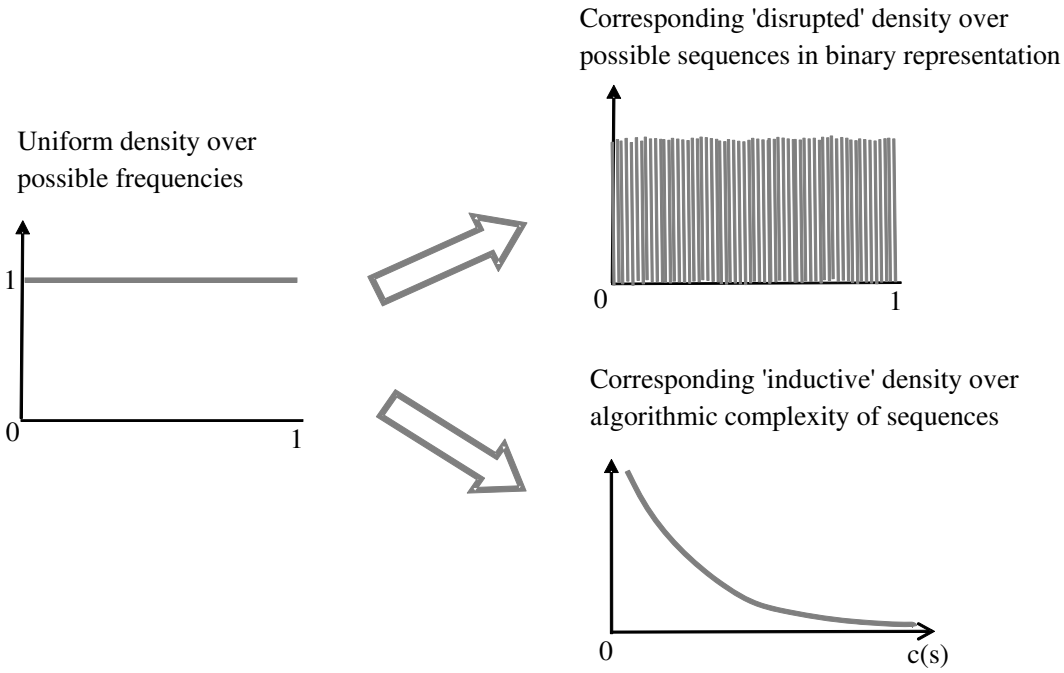


Figure 2. Transformation of a frequency-uniform into –

upper part: – a state-uniform density distribution.

lower part: – a distribution over the algorithmic complexity.

worlds in which meta-induction dominates random guessing is zero. Nevertheless many such worlds exist and it is precisely in these worlds that intelligent prediction methods can have chance at all. We should certainly not exclude these induction-friendly worlds from the start by assigning a probability of zero to them. This concludes my discussion of the relation between meta-induction and the no free lunch theorem within the perspective of the long run.

4. No Free Lunch and Meta-induction in the Short Run Perspective

The discussion of Wolpert's theorem within the perspective of the short run is more intricate. Recall that for finite sequences the advantage of meta-induction comes at a certain cost, that vanishes in the long run but is non-negligible for short sequences. Table 1 presents the result of a computer simulation of all possible prediction games with a length of 20 rounds, with binary events, three independent prediction methods and wMI.[4] The considered independent methods were

– majority induction, M-I, which always predicts the event that so far has been in the majority, and 0.5 in the case of ties (i.e.,$pred_{n+1}$ = 1/0.5/0 iff $f_n(1)$ >/=/< 0.5, respectively),

– majority anti-induction, M-AI, which predicts the opposite of M-I (i.e., $pred_{n+1}$ = 0/0.5/1 iff $f_n(1)$ >/=/< 0.5, respectively),

– averaging, Av, which always predicts 0.5.

Table 1 displays the frequencies of sequences for which the absolute success of a prediction method lies in a certain interval that is specified at the left margin, with [0,1) being the lowest and [19,20] the highest possible success interval. In accordance with the weak no free lunch theorem one sees in the bottom line that the average success is the same for all four methods. Nevertheless the frequency distributions over classes of sequences in which these methods reach certain success levels is remarkably different. The averaging method predicts always 0.5 and earns a sum-of-scores of 10 in all possible sequences. The object-inductive method M-I reaches a high success level in more worlds than the anti-

inductive method M-AI (symmetrically, Av-AI attains a low success level in more worlds than Av-I). In compensation, the number of worlds in which the anti-inductive method does just a little better than averaging is significantly higher than the corresponding number of worlds for the inductive method.

Sum-of-scores intervals	M-I	M-AI	Av	wMI
[0,1)	0	0.000	0	0
[1,2)	0	0.003	0	0
[2,3)	0	0.029	0	0
[3,4)	0	0.159	0	0
[4,5)	0	0.618	0	0
[5,6)	0.537	1.824	0	0
[6,7)	3.540	4.254	0	0
[7,8)	9.579	8.035	0	0
[8,9)	15.622	12.476	0	36.491
[9,10)	18.346	16.065	0	23.472
[10,11)	17.915	18.157	100.000	14.835
[11,12)	15.046	17.510	0	11.880
[12,13)	10.266	12.854	0	7.469
[13,14)	5.635	6.305	0	3.595
[14,15)	2.448	1.611	0	1.513
[15,16)	0.821	0.098	0	0.560
[16,17)	0.204	0	0	0.153
[17,18)	0.035	0	0	0.029
[18,19)	0.004	0	0	0.003
[19,20)	0	0	0	0
State-uniform average	10	10	10	10

Table 1. Computer simulation of M-I, M-AI, Av and wMI in all (2^{20}) binary sequences with 20 rounds. Cells show percentage of sequences in which certain levels of absolute success (left margin) have been reached.

Based on these results we obtain a justification of object-induction and of meta-

[4] Computer simulations were performed by Paul Thorn.

induction *even within* the induction-hostile perspective of a state-uniform prior distribution for *short-run* sequences. One can reasonably argue that what counts is to reach *high* success in those environments which *allow* for high success. This is what independent inductive methods do. At the same time one should *protect* oneself against *low* successes – this is what cautious methods of the type "averaging" do. The advantage of wMI meta-induction is that it combines *both* – reaching high successes where it is possible (inspect the intervals [12,13)–[19,20]) and at the same time avoiding low successes (inspect the intervals [8,9) and [9,10)). Thus wMI achieves "the best of both worlds". This, however, goes on the cost of a certain short-run loss (inspect the intervals [10,11) and [11,12)).

5. Conclusion

In this paper we confronted the optimality of meta-induction with the no free lunch theorem. We demonstrated that the apparent conflict between these two results disappears when one considers that the no free lunch theorem assumes a state-uniform prior distribution over the set of all (binary) event sequences. This distribution assigns a probability of zero to all infinite sequences that exhibit some sort of regularity which an intelligent prediction method could exploit. Short sequences were investigated by means of a computer simulation of all possible sequences of length 20. The result shows that in spite of having an equal expected predictive success, different prediction methods differ significantly in the frequency with which they reach certain success levels. Meta-induction turns out to offer the best combination of two abilities: exploiting regular sequences and avoiding loss-

es in irregular sequences.

We emphasize that this characterization of the advantage of meta-induction holds for the induction-hostile state-uniform prior distribution. If one switches to a frequency-uniform prior distribution, the computer simulation produces rather different results: Now M-I and wMI have highest predictive success in all classes of sequences whose frequencies are in the intervals [0,0.1), …, [0.3,0.4) and [0.6,0.7), …, [0.9,1]. wMI suffers from a small loss compared to M-I in these frequency intervals. In the frequency intervals [0.4,0.5) and [0.5,0.6) the picture is reversed: Here M-AI and Av are more successful than M-I; wMI suffers from a small loss compared to M-AI and Av, but is more successful than M-I. Because of space limitations we abstain from presenting the details.

References

Carnap, Rudolf. 1950. *Logical Foundations of Probability*. Chicago: Univ. of Chicago Press.

Cesa-Bianchi, Nicolo, and Lugosi, Gabor. 2006. *Prediction, Learning, and Games*. Cambridge: Cambridge Univ. Press.

Earman, John. 1992. *Bayes or Bust?* Cambridge/Mass.: MIT Press.

Howson, Colin, and Urbach, Peter. 1996. *Scientific Reasoning: The Bayesian Approach.* Chicago: Open Court (2nd ed.).

Reichenbach, Hans. 1949. *The Theory of Probability*. Berkeley: University of California

Press.

Skyrms, Brian. 1975. *Choice and Chance*. Encinco: Dickenson (4th ed. Wadsworth 2000).

Schurz, Gerhard. 2008. "The Meta-Inductivist's Winning Strategy in the Prediction Game: A New Approach to Hume's Problem."*Philosophy of Science* 75: 278-305.

Schurz, Gerhard, and Thorn, Paul. 2016. "The Revenge of Ecological Rationality: Strategy-Selection by Meta-Induction." 26(1), 2016, 31-59.

Solomonoff, Ray J. 1964. "A Formal Theory of Inductive Inference." *Information and Control* 7: 1-22 (part I), 224-254 (part II).

Wolpert, David H. 1996. "The Lack of A Priori Distinctions between Learning Algorithms." *Neural Computation* 8/7: 1341-1390.

Address of the author:
Professor Gerhard Schurz
DCLPS, Department of Philosophy
Heinrich Heine University Duesseldorf
Geb. 24.52, Universitaetsstrasse 1
40225 Duesseldorf, Germany
E-Mail: schurz@phil.uni-duesseldorf.de

Constructing Diagrams to Understand Phenomena and Mechanisms

Benjamin Sheredos
William Bechtel

Department of Philosophy
& Center for Circadian Biology
UC San Diego

Biologists often hypothesize mechanisms to explain phenomena. Our interest is how their understanding of the phenomena and mechanisms develops as they construct diagrams to communicate their claims. We present two case studies in which scientists integrate various data to create a single diagram to communicate their major conclusions in a research publication. In both cases, the history of revisions suggests that scientists' initial drafts encode biases and oversights that are only gradually overcome through prolonged, reflective re-design. To account for this, we suggest that scientists only develop a unitary understanding of their results *through* their attempts to communicate them.

1. Introduction

In biology, explanation often involves characterizing a phenomenon and generating an account of the *mechanism* thought to be responsible for it. The notion of mechanism has played this role in the life sciences since at least the 18th century, when it was adopted to characterize explanations that result from analyzing or decomposing biological systems into component parts, detailing their operations, and determining how these parts are organized and the operations orchestrated to produce the phenomena of interest (Bechtel & Richardson, 1993/2010). Scientists frequently find it productive to represent both phenomena and mechanisms in diagrams, in which different glyphs (Tversky, 2011) are laid out spatially. Shapes represent entities (the mechanism or its parts) and arrows represent operations. Space-on-the-page sometimes represents physical space (e.g., the nucleus versus the cytoplasm of the cell) but often is used simply to separate glyphs, distinguishing the represented parts and operations (Sheredos, Burnston, Abrahamsen, & Bechtel, 2013). Often, viewers can mentally animate a diagram to get an intuitive understanding of the mechanism's operations (Hegarty, 1992). Diagrams also aid in producing abstract mathematical cognition in the construction of computational models (Jones & Wolkenhauer, 2012).

Our focus here is on how scientists *generate* such diagrams. Generally these figures do not arise in a final format all at once, but result from a history of producing and revising interim drafts. Hand-drawn sketches might be preserved in laboratory notebooks (Nersessian, 2008). In the electronic era, a lineage of drafts is often preserved digitally in the files researchers save on the way to a final diagram. We take advantage of these to study the development of diagrams that appeared in two published papers. The first authors of each paper have provided their drafts of both text and the figures, allowing us to analyze their development. The last figure in one paper depicts a mechanism proposed to explain a phenomenon, whereas in the other the last figure presents a new phenomenon to be explained. We examine the drafts leading to these.

The researchers made major revisions to these diagrams as work on the manuscripts proceeded, reflecting the cognitive labor required in their development. We advance a hypothesis regarding why such labor is required, and why it takes the form it does: prior to the attempt to develop a coherent diagram, scientists themselves lack full understanding of the domain of inquiry. They have a “cognitive collage” (Tversky, 1993)

consisting of somewhat isolated and only partly-integrated understandings, often gleaned from diverse sources of data. It is through the attempt to develop a communicative diagram that these disparate, partial understandings are integrated into a detailed, cohesive whole.

Our case studies both concern research on circadian rhythms: endogenously-generated oscillations of approximately 24 hours that regulate the timing of other physiological and behavioral activities. The laboratory from which these publications arose studies circadian rhythms in cyanobacteria (specifically *Synechococcus elongatus*), the only bacterial lineage in which circadian rhythms have been demonstrated.

The basic mechanism responsible for circadian timekeeping in cyanobacteria is represented in Figure 1. (This figure comes from our first case study, examined further in the next section. For further details on the core mechanism see Kim, Dong, Carruthers, Golden, & LiWang, 2008.) The mechanism involves three proteins (KaiA, KaiB, and KaiC) plus their states and interactions at four major time-points (organized here in a circle, with different time-points at top, right, bottom, and left). KaiC is the large macromolecule shown at each time-point. It undergoes phosphorylation and dephosphorylation at two locations; the added phosphate groups are symbolized by the letter P in a black circle. KaiC itself initiates both phosphorylation and dephosphorylation, but the other two Kai proteins determine which dominates. When KaiA, represented using a purple, "bunny-eared" icon (top, right, and bottom), binds to KaiC (see top) phosphorylation is sped up and KaiC quickly becomes phosphorylated at both locations (see the two "P"s at right). When KaiB, represented using four stacked red ovals (at right and bottom) binds, it sequesters KaiA (see bottom), allowing dephosphorylation to proceed until neither location is phosphorylated (see left).

Since there is a specific and regular order of phosphorylation, and one cycle takes about 24 hours, KaiC's phosphorylation state predicts the current time of day, and serves as the cyanobacterium's "clock." Although open questions remain, this basic mechanism is well-established (see Mackey, Golden, & Ditty, 2011, for review) and provides the backdrop for the research pursued in our two case studies.

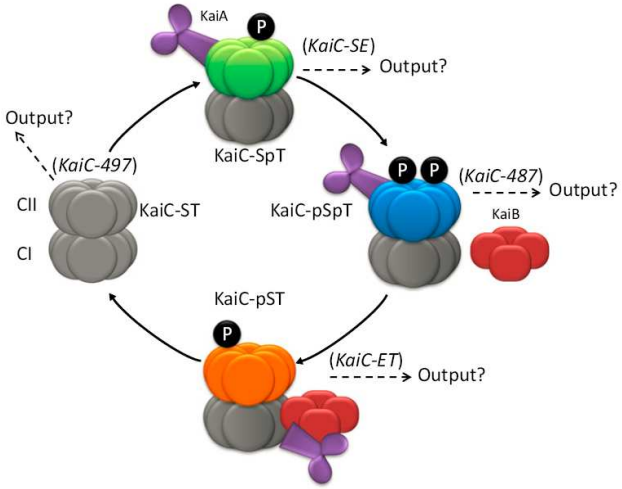


Figure 1. The first figure in Paddock et al. (2013) showing KaiC's phosphorylation cycle, as regulated by KaiA&B. Arrows leading to the word *Output?* encode uncertainty regarding which phosphorylation state communicates predicted time-of-day to the rest of the cell.

2. Advancing a New Hypothesis About the Output Mechanism

In our first case, Paddock, Boyd, Adin, and Golden (2013) advanced an important revision of what had become the standard account of the output mechanism through which the cyanobacterial circadian clock regulates the expression of virtually all genes. Two relatively well-defined classes of genes exhibit peak expression around (predicted) dawn and (predicted) dusk. If expression peaks near dawn, the gene is said to be regulated by a Class 1 promoter; if near dusk, by a Class 2 promoter. The proximal cause at work in each case is a transcription factor, which activates the promoter and initiates gene expression. Somehow, the clock must regulate promoter activation. Yet the Kai proteins are not transcription factors: none can directly influence any gene's expression. So additional components, forming an "output pathway," must mediate clock control of gene expression.

Two proteins, SasA and RpaA, had long been implicated since knocking out these proteins severely reduces rhythmic gene expression, even though the clock (KaiC's phosphorylation rhythm) is left intact. Since RpaA, but not SasA, is a transcription factor, the output pathway was hypothesized to run from KaiC to SasA to RpaA (Takai, Nakajima, Oyama, Kito, Sugita, Sugita, Kondo, & Iwasaki, 2006). We return to discuss this SasA-RpaA pathway below. What this research had not been able to determine, however, was which phosphorylation state(s) of KaiC triggers output.

This question is posed in Figure 1 (which is Paddock et al.'s published Figure 1). In addition to the glyphs we discussed above, the graphic includes four dotted arrows, originating at each phosphorylation state of KaiC and terminating in "*Output?*" Use of question marks to indicate uncertainties is common in mechanism diagrams. These arrows were added in a late draft (March 13, 2013), but the uncertainties they represent were formulated well in advance, as the specific target of research: Paddock et al. tested which phosphorylation state(s) drives output from the clock, and controls gene expression. The decisive experiments involved two steps.

First, Paddock et al. took cells and knocked out KaiC, destroying the clock. In this condition, there is no circadian *regulation* of gene expression: transcription factors activate promoters at their leisure. It was observed that with circadian regulation eradicated, Class 1 promoters "default" to a constantly high level of activation compared to wild-type (rather than selectively increasing activation at dawn) and Class 2 promoters "default" to constantly low activation (rather than selectively increasing activation at dusk).

Next, Paddock et al. reasoned that any phosphorylation state of KaiC that induced a deflection *away* from these "default" values could play some role in controlling output. To examine this, they created four molecules, each of which mimicked one phosphorylation state. (These phosphomimetics are named in italics in Figure 1). They then replaced KaiC with one of the phosphomimetics. Each modified cell essentially has a clock that is artificially "stopped" at one time-of-day. Paddock et al. then measured the effect on gene expression (using a luciferase reporter to detect promoter activation).

Only one phosphomimetic (*KaiC-ET*) induced activation different from the KaiC knockout's "default" activation. It both repressed the default-high activation of Class 1 promoters and enhanced the default-low activation of Class 2 promoters. This established that a single phosphorylation state of KaiC serves as the clock's output signal. This phosphorylation state (labeled "KaiC-pST" on the bottom of Fig. 1) corresponds, roughly, to predicted middle-of-the-night.

Between October 2012, when the authors began writing the manuscript, and June 2013, when they submitted it for publication, Paddock et al. drafted a series of diagrams to resolve the uncertainties presented in Figure 1. Two early versions are shown in Figure 2. Figure 2A continues to show all four phosphorylation states of KaiC, and adds an inhibitory arrow showing repression of P_{kaiBC} (a Class 1 promoter which serves to represent all Class 1 promoters) and an excitatory arrow showing activation of P_{purf} (representative of Class 2 promoters). Figure 2B partially simplifies the diagram by leaving out phosphorylation states that were ineffective in regulating gene expression, retaining a circle to indicate the phosphorylation cycle of KaiC. It also adds some linguistic labels and an indication that the clock is affected by inputs.

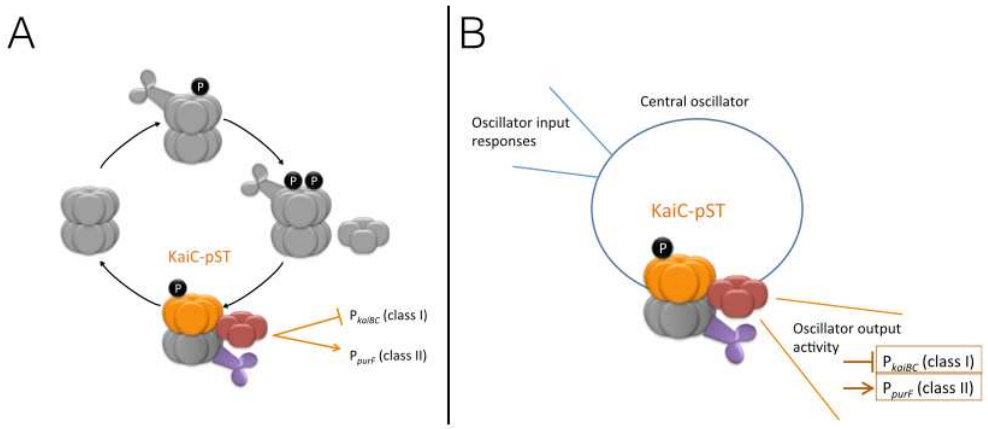


Figure 2. Panel A: an early version in the lineage of sketches that culminated in Figure 7, dated December 4, 2012. Panel B: a pared-down version, dated January 11, 2013.

These early drafts foregrounded the importance of KaiC-pST in regulating clock output, but did not include roles for SasA and RpaA, which were known to influence output. Paddock et al.'s data showed that when their phosphomimetic induced output from the clock, it did not affect the SasA-RpaA pathway. This left a puzzle: RpaA had been supposed to *mediate* the output, and yet it was not affected by the newly-identified output signal. To resolve this puzzle, the researchers drew upon additional data involving RpaA knockouts. As noted above, with KaiC knocked out (but with RpaA present), Class 1 promoters "default" to high activation, and Class 2 promoters "default" to low activation. In a RpaA knockout with KaiC still present these values are reversed: Class 1 promoters show constant low activation, and Class 2 promoters show constant high activation. When both RpaA and KaiC were knocked out, the results match those observed in the KaiC knockout alone. Taken together, the data indicate (a) that KaiC-pST affects output independently of RpaA, and (b) that the influence of RpaA is antagonistic to, or inhibitory of, the effects of KaiC-pST. Paddock et al. concluded that there were *two* output pathways: the previously-known SasA-RpaA pathway, and the one demonstrated to originate from KaiC-pST.

They drafted new diagrams (starting March 2013) to try to show how the two pathways interact in regulating gene expression. Presumably because they wanted to show the origin of the SasA-RpaA pathway in a different phosphorylation state, they restored the

other phosphorylation states that had been dropped from Figure 2B (see Figure 3A below). Because the origin of the SasA-RpaA pathway was unknown, the sketch shown in Figure 3A does not link it to any one phosphorylation state. It is shown as inhibiting output from the KaiC-ST phosphoform. Here effects on gene expression are shown all at once, in terms of a general measure, Oscillator Output Activity, which we do not discuss further (an analysis has been provided by Burnston, Sheredos, Abrahamsen, and Bechtel, 2014).

Altogether Paddock et al. generated seven variants of Figure 3A. These became quite complex as they tried to illustrate the interactions between pathways. Then in the draft of April 11 they abruptly changed to the simpler format shown in Figure 3B. Multiple representations of KaiC's phosphorylation states are removed, and a single circle represents KaiC's phosphorylation rhythm. Another circle represents a cycle of RpaA phosphorylation. Instead of trying to show the effects on Class 1 and Class 2 promoters at once, they show each separately, duplicating the whole arrangement. The schematic graphs on the right indicate effects on expression (using recorded bioluminescence as well as the measure of Oscillator Output Activity, now renamed Kai Oscillator Activity (KOA)).

Figure 3B shows what appeared, with minor changes, as part of Paddock et al.'s Figure 7. The history of drafting described here reflects a variety of attempts to portray the mechanisms of circadian output. What we highlight is the progression through several repetitive phases of abstraction, or the elimination of detail. In moving from Figure 2A to 2B, a number of details regarding phosphorylation states of KaiC are deemed irrelevant and dropped out. Yet when it comes time to add a depiction of the RpaA pathway, these same details re-appear in Figure 3A. Along with a number of other changes, there is a repetition of the same abstraction to obtain 3B, and the same details are again dropped out. This months-long drafting process only gradually produced the published figure, and one sees the researchers struggling repeatedly to move away from their initial, detail-rich sketch.

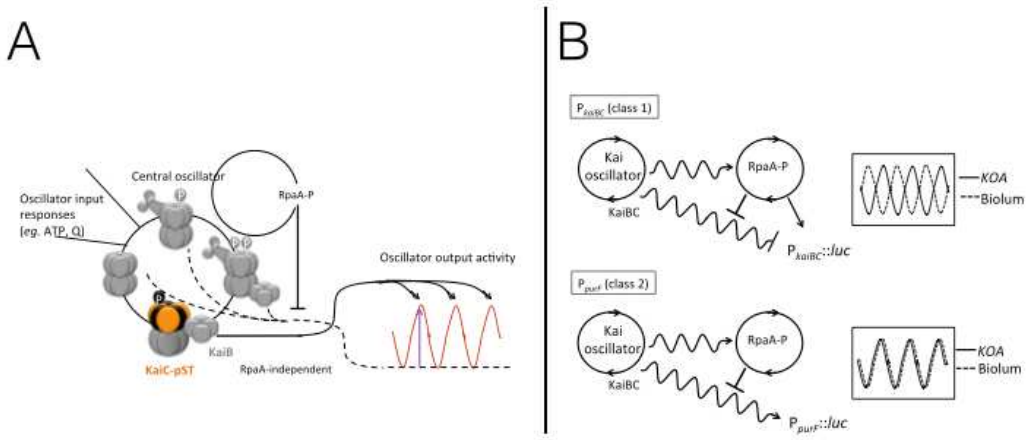


Figure 3. Panel A: one of several intermediate versions that appeared in versions of the manuscript from early March until mid-April 2013. Panel B: a pared down version that appeared on April 11, 2013.

3. Characterizing the Changing Location of the Clock within the Cell

Our second case comes from the same laboratory. Instead of advancing a new set of operations in a mechanism, Cohen, Erb, Selimkhanov, Dong, Hasty, Pogliano, and Golden (2014) reveal a new phenomenal aspect of the circadian clock, its changing location within the cell over the course of a day. Although the clock's migration is potentially important in explaining the operation of the clock, the goal of the paper is simply to demonstrate this movement.

Since they lack internal membranes, bacteria were long regarded as internally disorganized bags of genes, enzymes, and other molecules. Recent research has identified extensive internal organization and determined its importance for various physiological activities of bacteria (Rudner & Losick, 2010). Cohen et al. set out to investigate where the Kai proteins are located in the cell. Using luciferase and fluorescence reporters, they determined that although KaiA and KaiC are distributed throughout the cell during the day, at night they localize to one pole. Notably, when KaiA and KaiC are localized at the cell pole, they are co-localized with KaiB and with CikA (a part of the input pathway to the clock, affecting its "entrainment" or synchronization to local day/night cycles). Cohen et al. suggest that this localization may be functionally significant for timekeeping, and may "facilitate interactions among the clock components" (p. 1840).

The data graphics for the paper, presenting evidence for the changing localization of KaiA and KaiC, were largely settled by the time drafting of the manuscript began in March 2014. Over the following four months of drafting, much effort was spent developing a diagram linking the localization of the proteins to previously-known operations involved in the clock. All versions were prepared by the first author, with others offering advice and aiding decisions between alternatives. We examine a few steps between the initial draft and the final figure, which eventually appeared as the final figure in the published article.

The initial draft, dated March 7, 2014 comprised five panels. Three panels reproduced extant images (from the web or from another publication) to present some examples of how other diagrams had shown relevant information. One panel consisted of questions and design considerations for the figure, and read:

"Model: entrainment/proteolysis.

1. SDH/respiration goes to poles in low light.
2. ATPase interactions at night? ATPases are in the curved part of the chloroplasts
3. Curvature"

Although these phrases are cryptic, they reference specific information that the researchers considered including in their diagram. The word *entrainment* refers to CikA's role in the input pathway. The co-localization of CikA and KaiC at the cell pole may be related to entrainment, and it was considered whether to emphasize this in a future draft. The word *proteolysis* invokes a well-documented migration of proteins to the membrane when targeted for destruction. The newly-documented movement of the Kai proteins was found not to be related to proteolysis, and it was considered that future sketches might underscore this. The first bullet point points out that succinate dehydrogenase (SDH) and other enzymes involved in respiration also migrate to the poles. Implicitly, the question of the relation of the clock's migration to basic cell metabolism is being raised. The next bullet point raises it more explicitly, asking whether the clock's migration is related to energetic

processes at the pole. The last bullet point raises the question of whether the curvature of the membrane at the pole figures in directing the migration.

Although these were raised as design considerations, none were addressed in the initial draft that appeared in the remaining panel (Figure 4 below). A single cyanobacterium is shown with a green line representing its membrane. To illustrate the different state of the clock over time, the figure is divided diagonally into two segments (day phase is yellow, including an icon of the sun, and night is grey, including a moon). In each half of the figure, the phosphorylation cycle is shown, using glyphs similar to Paddock et al.'s for the Kai proteins, but showing only two of the four phosphoforms (white circles indicate phosphates). During the day phase KaiA is shown bound to phosphorylated KaiC, and detached from unphosphorylated KaiC; all these glyphs are situated towards the center and away from the pole. In the night phase CikA and KaiB are included, and the glyphs are placed near the pole. Two bullet points reference other studies documenting events occurring during the night phase.

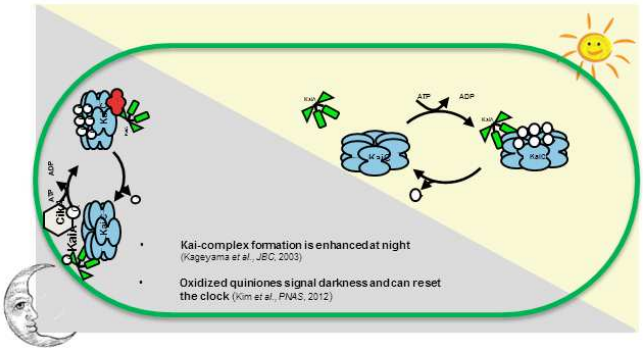


Figure 4. First draft of the mechanism diagram in Cohen et al. Dated March 7, 2014.

This first draft did not address any of the additional design issues discussed above (entrainment, proteolysis, metabolism, etc.). Based on feedback the first author received, she prepared two revised versions (Figure 5 below). Neither addressed those additional design issues but instead focused on the phenomenon of localization alone. Two features shared by these new drafts are the use of a vertical rather than diagonal division of the figure into sections for day and night, and the incorporation of a diamond representation in the center for the full, four-stage cycle of KaiC phosphorylation.

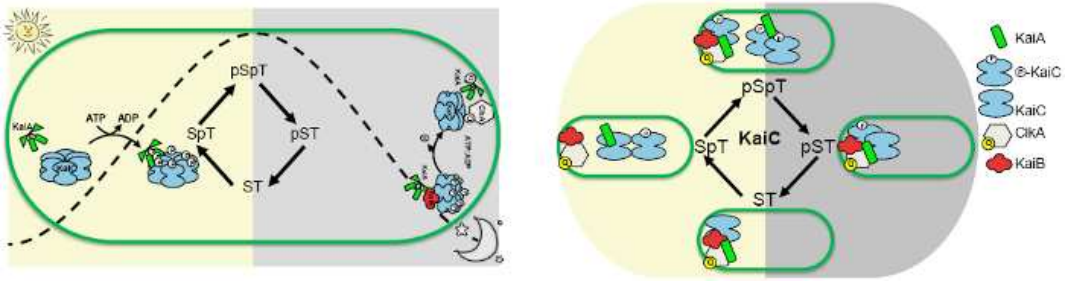


Figure 5. Two drafts of Cohen et al. Figure 6 from April 6, 2014.

The version in Figure 5's left panel retains the portrayal of one bacterial cell, with the Kai complex localized to the pole at night (now on the right) and free in the cytoplasm during the day. Overlaid is a bell-curve representation of the abundance of KaiC, which increases in concentration during the day and declines through the night. The implied x-axis for this graph imposes a linear representation of time, from dawn on the left to the next day's dawn on the right. This is in tension with the cyclical representation of time in the center of the figure. This kind of infelicity is not surprising in a draft diagram, as theauthor is actively trying out ideas in the attempt to construct a coherent representation. Despite this infelicity, the first author prefers this version, and continues to use it in her talks.

Other members of the research team, however, preferred the version in Figure 5's right panel, which introduces a fundamental change: multiple representations of the cell, aligned with the four phosphorylation states of KaiC, shown in the cyclical representations in the center. It is interesting that the whole figure takes the form of an oval although there is no longer any attempt to show all processes within a single bacterial cell: this is a "remnant" from the first sketch. Finally, this figure introduces a legend to link different glyphs to the molecules they represent.

After feedback from the other authors, the first author created two more versions by April 25, 2014. One of these (Figure 6 below) was eventually published without any further alteration. The sun and moon glyphs are re-introduced, and the spacing of some protein glyphs is slightly altered. Perhaps the most significant innovation is that additional KaiC icons in light blue are added in all stages of the cycle. This is intended to indicate that there are many copies of KaiC in the cell, and they are often in different states of phosphorylation.

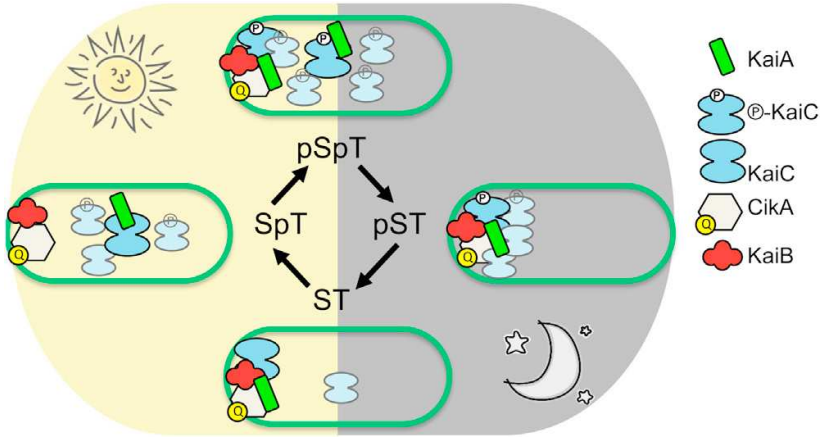


Figure 6. Final version of Figure 6 in Cohen et al.

From the initial sketch, the diagram underwent substantial modification until the authors settled on the published version. We highlight that the final version includes much *more* detail than was present in the initial sketch, and that the history of revision is one of gradually adding details. The first step was to consider a variety of details that had been omitted (e.g., regarding entrainment, proteolysis, metabolism). These were not addressed; instead the authors added a clockwise portrayal of the progression of phosphorylation

states, and exploited this to organize the newly-discovered information regarding Kai localization at different times of day. The fundamental constraint was to deploy limited space-on-the-page to simultaneously represent intracellular space, functional states of the components, and time-of-day. This proved to be difficult: several rounds of revision were required before a format was attained which overcame the limitations of the initial sketch. Even still, many graphical elements changed little through the revisions.

4. An Hypothesis: Scientists Develop Understanding by Drafting Diagrams

We examined the evolution of two diagrams developed to communicate hypotheses about phenomena and mechanisms. Through analysis of the lineage of drafts the authors made, we identified an iterative process in which different representational strategies were gradually developed and enacted. In one case, the initial sketch was much more detailed than the final graphic; the excess details stubbornly re-appeared mid-way through revisions, only to be dropped again, revealing an iterative attempt to pare down irrelevant detail. In the other case, the initial sketch required supplementation to reach the desired degree of detail, and basic limitations of the initial sketch had to be gradually overcome in order for details to be coherently included. We described an iterative attempt to add in relevant detail.

One might regard the development of these diagrams as essentially epiphenomenal to scientific cognition: at the outset, the researchers possessed a cohesive understanding of the domain, and diagram construction was an additional layer of practice, aimed at developing a representation to aid in *communicating* that pre-established understanding. We question this "epiphenomenalist" view. The histories of revision suggest that at the outset it was neither obvious to scientists what details should be included in an adequate diagram, nor obvious how relevant details might be adequately represented. Rather, an initial attempt was made, and its excesses and omissions were *then* identified and corrected. The epiphenomenalist might account for this by proposing some general cognitive inability to communicate the cohesive understanding of the domain which researchers allegedly had in advance. But construction of these diagrams was preceded by months of reflective and careful experimental work, resulting in a hard-won understanding of the domain that motivated the researchers to write a manuscript in the first place. We can agree with the epiphenomenalist that translating the pre-established understanding into a specifically *graphical* format is an important challenge. But this is not the whole story. First, the data, which support the understanding of the domain, are typically *already* encoded in a graphical format – in the data graphics which, in our cases, were essentially finalized before authors begin the months-long process of developing their diagrams. (Moreover, both our cases show researchers integrating data-graphics directly into mechanism diagrams, suggesting there is little if any cognitive "gap" between them.) Second, while it takes multiple revisions to generate a diagram which is deemed "just right," there is little reason to posit any inability to develop graphical representations as such: witness the number and variety of graphics the researchers developed, of which we have given only a small sample.

An adequate account must grant that researchers began with *some* understanding of the hypotheses they aim to communicate, but must account for the great expenditure of cognitive labor documented in the history of revisions. We propose that as a result of experimental work, scientists understand the domain through a variegated "cognitive

collage" (Tversky 1993) involving a diversity of representational formats. Some might be abstract (e.g., mathematical), others more clearly materially grounded (e.g., embodied familiarity with experimental protocols), and most are probably a mix. These representations are sufficiently integrated to enable the researcher to articulate their major hypothesis, but they are not yet integrated in a single representation that simultaneously provides an adequate understanding of the evidence for, and relations between, various elements of the hypothesis. There is at this point no "map-like" representation that integrates all this information. The initial sketches, we propose, are the first attempt to integrate this information into a cohesive representation. This integrative process is prone to what scientists regard as errors, of which we have identified two kinds: the inclusion of irrelevant and the omission of relevant detail. Moreover, the process is prone to a kind of anchoring effect—the initial sketch may include infelicitous elements that persist as an obstacle for later revisions (e.g., the re-appearance of irrelevant detail in Paddock et al.'s revisions; the difficulty of representing time in Cohen et al.'s graphics).

A variety of authors have argued for the practical necessity and epistemic merits of "multiple models idealization" (see Weisberg 2007 for references to the many proponents of this view). Viewed in the context of that work, our hypothesis provides an account of how and when such integration can be achieved: an iterative process of diagram redesign can generate the final, cohesive understanding of the phenomena or mechanism.

References

Bechtel, W., & Richardson, R. C. (1993/2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.

Burnston, D. C., Sheredos, B., Abrahamsen, A., & Bechtel, W. (2014). Scientists' use of diagrams in developing mechanistic explanations: A case study from chronobiology. *Pragmatics and Cognition*.

Cohen, S. E., Erb, M. L., Selimkhanov, J., Dong, G., Hasty, J., Pogliano, J., & Golden, S. S. (2014). Dynamic localization of the cyanobacterial circadian clock proteins. *Current Biology, 24*, 1836-1844.

Hegarty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 1084-1102.

Jones, N., & Wolkenhauer, O. (2012). Diagrams as locality aids for explanation and model construction in cell biology. *Biology and Philosophy, 27*, 705-721.

Kim, Y.-I., Dong, G., Carruthers, C. W., Golden, S. S., & LiWang, A. (2008). The day/night switch in KaiC, a central oscillator component of the circadian clock of cyanobacteria. *Proceedings of the National Academy of Sciences, 105*, 12825-12830.

Mackey, S. R., Golden, S. S., & Ditty, J. L. (2011). The itty-bitty time machine: Genetics of the cyanobacterial circadian clock. *Advances in genetics, 74*, 13-53.

Nersessian, N. (2008). *Creating scientific concepts*. Cambridge, MA: MIT Press.

Paddock, M. L., Boyd, J. S., Adin, D. M., & Golden, S. S. (2013). Active output state of the Synechococcus Kai circadian oscillator. *Proceedings of the National Academy of Sciences, 110*, E3849-E3857.

Rudner, D. Z., & Losick, R. (2010). Protein subcellular localization in bacteria. *Cold Spring Harbor Perspectives on Biology, 2*, a000307.

Sheredos, B., Burnston, D., Abrahamsen, A., & Bechtel, W. (2013). Why do biologists use so many diagrams? *Philosophy of Science, 80*, 931-944.

Takai, N., Nakajima, M., Oyama, T., Kito, R., Sugita, C., Sugita, M., Kondo, T., & Iwasaki, H. (2006). A KaiC-associating SasA–RpaA two-component regulatory system as a major circadian timing mediator in cyanobacteria. *Proceedings of the National Academy of Sciences, 103*, 12109-12114.

Tversky, B. (1993) "Cognitive maps, cognitive collages, and mental spatial models" *Spatial Information Theory a Theoretical Basis for GIS*, volume 716 of the series *Lecture notes in computer science*, 14-14.

Tversky, B. (2011). Visualizing thought. *Topics in Cognitive Science, 3*, 499-535.

Weisberg, M. (2007) Three kinds of Idealization. *J.Phil* 104(12), 639-659.

Bayesian Statistical Inference and Approximate Truth

Olav B. Vassend

October 28, 2016

Abstract

Scientists and Bayesian statisticians often study hypotheses that they know to be false. This creates an interpretive problem because the Bayesian probability of a hypothesis is supposed to represent the probability that the hypothesis is true. I investigate whether Bayesianism can accommodate the idea that false hypotheses are sometimes approximately true or that some hypotheses or models can be closer to the truth than others. I argue that the idea that some hypotheses are approximately true in an absolute sense is hard to square with Bayesianism, but that the notion that some hypotheses are comparatively closer to the truth than others can be made compatible with Bayesianism, and that this provides an adequate and potentially useful solution to the interpretive problem. Finally, I compare my "verisimilitude" solution to the interpretive problem with a "counterfactual" solution recently proposed by Jan Sprenger.

Contents

1 Introduction

According to the standard Bayesian interpretation of probability, the probability of a hypothesis is the probability that the hypothesis is *true*. However, scientists, including scientists who make use of Bayesian statistical methods, often investigate models and hypotheses that they know to be false. In particular, statistical models tend to be constructed on the basis of auxiliary assumptions (e.g. normality and independence of measurement errors) that are often known to be false. Moreover, statistical analysis is often restricted to hypothesis sets, such as the set of linear or exponential functional relationships, that are known to at best be (false) approximations of the actual functional relationships. Presumably, if something is known to be false, then it has a probability of 0 of being true, so all of the preceding practices are hard to reconcile with the standard Bayesian interpretation of probability. Indeed, Bayesian statistical practice apparently is faced with an interpretive problem: on

the one hand, Bayesian probabilities are standardly interpreted as probabilities of truth; on the other hand, Bayesian scientists routinely assign non-zero probabilities to hypotheses they know to be false.

How serious is the interpretive problem and how may it be solved? I argue that there are many cases where the interpretative problem does not arise, even when the statistical model is false. But there are also many cases where the interpretive problem does arise. Many scientific realists have suggested that successful scientific models and hypotheses, though usually false, are nonetheless often approximately true, or – at the very least – that successful hypotheses in general are "closer to the truth" (or have higher "verisimilitude") than hypotheses that are less successful. I argue that, provided we jettison the standard Bayesian interpretation of the probability axioms, Bayesianism can accommodate the insight that some false hypotheses are closer to the truth than others, and that this reinterpretation of the probability axioms is potentially useful. I contrast this solution to the interpretive problem with another recent proposal due to Jan Sprenger (2016), according to which probabilities of false hypotheses are interpreted as "counterfactual degrees of belief," and I argue that the two approaches – when spelled out in detail – are formally inter-translatable and help illuminate each other.

2 The Basics of Standard Bayesian Inference

Bayesianism is a prominent approach in both confirmation theory and in statistical inference. Bayesian confirmation theory and Bayesian statistics clearly have many things in common, but they are also different enough that it pays to discuss them separately. In this paper, I will focus my attention on Bayesian statistical inference, though much of what I will say also has relevance to Bayesian confirmation theory.

In statistical inference, a set of competing hypotheses is usually indexed by a *parameter*, which in general will be a real-valued variable or a vector of real-valued variables. Given a space of candidate hypotheses parameterized by Θ, and given some particular context in which the possible observations or outcomes are x_1, x_2, etc. – or X, for short – a *statistical model* consists of a set of conditional probability

(density) distributions, $p(x|\theta)$, that jointly specify the probability of each possible $x \in X$ given each possible $\theta \in \Theta$.[1]

Almost invariably, the statistical model is premised on various auxiliary assumptions, A, that jointly guarantee that each value of θ *entails* a probability for each x. Sometimes A itself has free parameters – so-called "nuisance parameters," N – that must also be estimated from the data, in which case the conditional probability distributions will be of the form $p(x|\theta\&n)$. Thus, a statistical model may in general be regarded as being composed of two distinct ingredients: the hypotheses of interest, parameterized by Θ; and the auxiliary assumptions, A, consisting of nuisance parameters, N, and background assumptions, B. It follows that a statistical model is "true" if and only if the following conjunction is true: (1) some element of Θ is true and (2) A is true: that is, B is true and some element of N is true.

For example, suppose you are interested in estimating the mass of some object by measuring it a single time using a scale. The hypotheses of interest are the various possible masses of the object, which you may index using a real-valued parameter, m. The possible outcomes, x, are the various possible outcomes of the measurement. In order to probabilistically link m to x, you may, for example, add the auxiliary assumption, A, that the measurement outcome is normally distributed around the true mass with a variance of d. Here d is a nuisance parameter. Then the assumptions of the statistical model generate the following conditional probabilities:

$$p(x|m\&d) = \frac{1}{d\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-m)^2}{2d^2}} \tag{2.1}$$

In this case, the statistical model is true if and only if (1) there is some value m_0 of m that corresponds to the actual mass of the object, and (2) the measurement outcome is actually normally distributed around m_0 with some variance d_0.

What distinguishes Bayesian inference from other sorts of statistical inference is that Bayesians use probability distributions to assess the plausibility of *parameter values*. In addition to requiring a statistical model, a Bayesian analysis therefore

[1]Note: p is a *probability function* over the set $\mathbf{X}$ if and only if the following three axioms are satisfied: (1) $p(\mathbf{X}) = 1$. (2) $p(x_i) \geq 0$ for all $x_i \in \mathbf{X}$. (3) $p(\bigvee x_i) = \sum p(x_i)$, whenever the x_i in the disjunction are mutually exclusive.

requires that the parameters of interest Θ and the nuisance parameters N all be assigned so-called *prior* probabilities; these are probabilities that are assigned before the observation of data. Moreover, if there are multiple candidate statistical models, then all of the models must be assigned prior probabilities as well. In the above example, prior probabilities must therefore be assigned to each possible value of m and to each possible value of d. Once these probabilities have been assigned, the *joint* distribution of the possible observations and the parameters is defined as the product of the likelihood and the prior: $p(x, m, d) = p(x|m, d) * p(m, d)$. The *posterior* probability distribution of m is given by Bayes's theorem, $p(m|x, d) = p(x|m, d) * p(m, d)/p(x, d)$

There is disagreement among Bayesians concerning how prior and posterior probabilities should be interpreted. Some see these probabilities as the subjective or rational degrees of belief of some agent, whereas others interpret them as evidential degrees of support or as representing an objective state of information. However, regardless of whichever more specific interpretation they endorse, Bayesians of all kinds agree that $p(\theta)$ represents the probability that θ is *true*.[2] This interpretation of probability – the standard Bayesian interpretation – leads to problems, however, because the models and hypotheses that scientists investigate are often believed or even *known* not to be true. This problem has not gone completely unnoticed in the philosophical literature,[3] but in general the seriousness of the problem seems not to have been appreciated. The problem seems to be more acknowledged in the statistical literature, but no satisfactory resolution has been offered.

3 The Interpretive Problem in Bayesian Statistical Inference

Statistical models, much like other models in science, contain idealizations and approximations that render the models strictly speaking false. Typical examples in-

[2]Or more precisely, the probability that the hypothesis indexed by θ is true.

[3]E.g. Forster and Sober (1994), Shaffer (2001), and more recently Sprenger (2016).

clude, e.g., the assumption that measurement error is bell-shaped, or that measurements are independent and identically distributed. To be sure, these assumptions are often justified because they hold *approximately*, but they rarely hold *exactly*. In other words, the auxiliary assumptions of statistical models are generally false.

For these reasons, the statistician George Box famously said that "all models are false, but some models are useful."[4] More recently, Andrew Gelman and Cosma Shalizi write, "To reiterate, it is hard to claim that the prior distributions used in applied work represent statisticians' states of knowledge and belief before examining their data, if only because most statisticians do not believe their models are true, so their prior degree of belief in all of Θ is not 1 but 0." (Gelman and Shalizi, 2013, p. 19).

A fully Bayesian analysis requires that we assign probabilities to our models and to the parameters inside the models. But according to the standard Bayesian interpretation of probability, the probabilities we assign are supposed to represent the probabilities that the models and parameters are true. If we know that they are all false, it would seem they should therefore be assigned a probability of 0.

Of course, Bayesian statisticians typically do not assign probabilities of 0 to parameters or to models; they assign non-zero probabilities. This practice is what leads to the interpretive problem, which may be phrased in the form of a question: what does it mean to assign a model or hypothesis that is known to be false a non-zero probability? To more precisely diagnose the problem, it helps to state the probability axioms with the standard Bayesian interpretation made explicit:

> Suppose $\mathbf{H}$ is a set of hypotheses $\{H_1, H_2, \ldots, H_n\}$. Then
>
> 1S. $p(\mathbf{H}) = 1$. Interpretation: one of the hypotheses in $\mathbf{H}$ is true.
>
> 2S. $p(H_i) \geq 0$ for all $H_i \in \mathbf{H}$. Interpretation: no hypothesis has a negative probability of being true.
>
> 3S. $p(\bigvee H_i) = \sum p(H_i)$, whenever it is impossible for more than one H_i in the disjunction $\bigvee H_i$ to be true.

[4] This quote is famous enough that it has a Wikipedia page. Box repeated the quote, or variations of it, in several places. e.g. Box (1980)

Here we can see that the interpretive problem is really a problem with the standard interpretation of the first probability axiom. That is, for many of the hypothesis sets that scientists study, it will not be the case that one of the hypotheses is true. Hence, strictly speaking, many hypothesis sets will not satisfy axiom 1S. Axioms 2S and 3S, on the other hand, will generally be satisfied by the kinds of hypothesis sets that Bayesian statisticians study.

One possible remedy to the interpretive problem that might initially seem attractice is to try to change the algebra over which the probability function p ranges.[5] Later, we shall consider a couple of specific proposals along these lines. However, there is a fundamental reason why any such proposal will not work. Briefly, the reason is that if you want to do Bayesian inference on a statistical model that is parameterized by θ, then you need to assign probabilities to θ; you cannot instead assign probabilities to, e.g., propositions of the sort $<\theta = 2$ is the best parameter value$>$ or $<\theta = 2$ is the parameter value that is most predictively accurate$>$, because these propositions are not part of the statistical model. Nor can you amend the statistical model so that it is instead parameterized by these other propositions.

Gelman and Shalizi's (2013) solution to the interpretive problem (to the extent that they see it as a problem) seems to be to refuse to interpret Bayesian probabilities in any standard way. Bayesian probabilities of parameters inside models, they say, are "regularization devices" and models themselves should not really be assigned probabilities at all. This does not seem like a solution so much as an admission of defeat. Morey et al. (2013) pursue a different strategy. They reply to Gelman and Shalizi with the assertion that "...scientific models, including statistical models, are neither true nor false" (p. 71) and that "Box's (1979) famous dictum... ...could be shortened to 'some models are useful' without any loss" (p. 71). They then recommend assigning odds rather than probabilities to models because a "Bayesian who employs odds is silent on whether or not she is in possession of the true model, and, in fact, need not acknowledge the existence of a true model at all" (p. 71). It is, however, unclear how using odds rather than probabilities is supposed to solve

[5]For example, some might be tempted to consider the algebra generated by the associated propositions, $<H_i$ is the best hypothesis$>$, for each H_i, or something similar.

the interpretive problem. And it is not clear how refusing to assign truth values to models solves the problem either. What does it mean to say that your odds are 5 to 1 in a model that is neither true nor false as against another model that is also neither true nor false? The interpretive problem seems to be just as severe here as before.

Moreover, the claim that statistical models do not have truth values seems wrong. As we saw, a statistical model can be regarded as a conjunction of a claim about the hypotheses of interest (namely that one of them is true) and a claim about the auxiliary assumptions (namely that they are all true). It follows that a statistical model is false either if none of the hypotheses of interest is true or if one of the auxiliary assumptions is false. The second situation arguably is less serious than the first.

4 False Auxiliary Assumptions vs False Hypotheses of Interest

If a statistical model is false because one of its auxiliary assumptions is false – which is almost always the case – then the interpretive problem arises on the level of model inference. That is, if there are multiple statistical models that all contain known false auxiliary assumptions, then all of the models will have a probability of 0 of being true, and hence a standard Bayesian who wants to use Bayesian inference to find the best model will run into the problem of how to sensibly assign non-zero probabilities to the models.

How to make sense of model inference and model selection is therefore a serious problem for Bayesians. However, if the statistical model is not itself the hypothesis of interest, then the fact that the statistical model is false does not necessarily mean we are faced with the interpretive problem.

Consider, for example, the previous example involving the estimation of the mass m of some object. A good way of getting an estimate of m is by embedding m in a statistical model. Now, even if the statistical model is false because it is based

on known false (auxiliary) assumptions, probabilistic statements about m will still be completely sensible; thus, in cases like this one, the interpretive problem does not arise for inferences about the parameter m. For example, a statement like "the probability that it's true that m is 2kg is 0.5" is perfectly sensible as long as it is remembered that the probability is premised on the auxiliary assumptions of the model. If those assumptions are seriously wrong, the probability may well be inaccurate or misleading; however, the probability can still sensibly be interpreted as a probability of truth.

Because Bayesian parameter inference often makes sense even if the statistical model is false, George Box famously recommended a reconciliation between Bayesian and frequentism. According to Box, frequentist methods should be used to identify a "useful" (albeit false) statistical model; Bayesian inference can then be used to infer plausible parameter values inside the assumed statistical model. This two-step procedure makes sense in cases where the hypotheses of interest are parameters that represent real quantities out in the world, such as for example the mass of an object.

However, it happens not infrequently in science that the hypotheses of interest are themselves known to be false, strictly speaking; but this has not stopped scientists from employing Bayesian methods in their research. For example, phylogeneticists in both biology and linguistics use trees to represent family relationships between species or between languages. In both cases, the trees investigated omit known relationships and introduce false idealizations. For example, a tree phylogeny for a language family is premised on the (false) idea that languages bifurcate instantaneously and are forever separated thereafter. Yet, even though all phylogenetic trees are clearly false, Bayesian phylogeneticists are often interested in discovering which tree has the highest posterior probability. These probabilities cannot comfortably be interpreted as probabilities that the trees are literally true, and thus we are faced with the interpretive problem.

The interpretive problem also arises whenever the hypotheses under consideration posit simple functional relationships that are almost certainly false idealizations. This is usually the case whenever Bayesian linear regression is used, for example, because most functional relationships in the world are not actually linear.

As an example, suppose you are interested in the functional relationship between just two variables, X and Y. For concreteness, suppose X represents some measurement of a complex system, e.g. the barometric pressure of a weather system, and Y represents some quantity of interest, e.g. how much it will rain in the next hour. The true functional dependence of Y on X is in all likelihood very complex.[6] Nonetheless, it is very common in such cases to restrict attention to classes of simple functional relationships, such as the set of linear hypotheses with 0 intercept, which models the relationship between Y and X as follows:

$$Y = \alpha X + \epsilon \tag{4.1}$$

Here, ϵ represents the (hypothesized) random fluctuation around the linear function $Y = \alpha X$; ϵ is generally taken to be a normal distribution with a mean of 0 and standard deviation d. α is the parameter of interest while d is a nuisance parameter (auxiliary assumption); both need to be estimated from data. Note that α does not represent some "real" quantity out there in the world; indeed, if we were to interpret α as representing a real quantity, then presumably that quantity would be a rate. Thus, α would refer to the constant rate at which Y changes (on average) given changes in X. However, if the true functional relationship between X and Y is not actually linear, then there is no constant rate at which Y changes in response to changes in X. Thus, in sharp contrast to the previous example concerning the estimation of the mass of an object, $\alpha = 2$ cannot be true or false in the same way that statements such as $m = 2$ or $m = 3$ are true or false.

But if α does not represent a quantity in the world, then what does it mean for a given value of α to be "true" or "false"? Well, α indexes a set of hypotheses, namely $Y = \alpha X + \epsilon$, so to say that α_0 is "true" in this case is the same as saying that there *exists* some value of ϵ such that the hypothesis $Y = \alpha_o X + \epsilon$ is the true functional relationship between X and Y. To paraphrase Sober (2015), α "lives inside" its

[6]By "the true functional dependence," I mean the functional dependence that would result if we were to keep fixed all other predictively relevant variables and see how Y varies given changes in X. Since X may – indeed probably does – interact with other variables, this definition is too simplistic, but going into the details here is not worth the pay-off.

model; it has a meaning only in the context of the statistical model of which it is a part. Not all parameters are created equal.

Note that in this example, it is pretty much a foregone conclusion that *no* hypothesis of the form $Y = \alpha X + \epsilon$ describes the true functional relationship between Y (i.e. how much it will rain) and X (the barometric pressure). Hence it's not merely the auxiliary assumptions of the model that are false in this case; the very hypotheses that we are interested in are all known in advance to be false. Hence, the interpretive problem hits us again with full force: how are we supposed to understand non-zero probability assignments to values of α?

5 Approximate Truth

This is where the notion of approximate truth may be helpful. More generally, scientific realists would doubt whether any scientific or statistical model could be "useful" (to use Box's term) were it not approximately true in some sense; thus, we should assign a model (or a parameter inside a model) a probability proportional to the extent to which we find it approximately true (in the relevant sense). The question we need to ask is whether and how the idea that hypotheses and models are sometimes approximately true or that some hypotheses are closer to the truth than others can be accommodated within the Bayesian framework. Because model inference and parameter inference are different in some important ways, I will from now on focus only on parameter inference. That is, I will assume that the hypotheses of interest are indexed by a parameter Θ inside some fixed statistical model, and that each $\theta \in \Theta$ picks out some hypothesis that does not itself contain adjustable parameters.

Before we can address properly the question whether some hypotheses can be approximately true or closer to the truth than others, we must make a few assumptions about what approximate truth is and how it can be measured.

The study of approximate truth was initiated by Popper (1963) and has by now

accumulated a large literature.[7] The most influential contemporary approach in the study of approximate truth – known in the literature as the "similarity approach" – takes seriously the idea that approximate truth is a particular kind of approximation. To say that something is a good approximation of something else is to say that the two things are similar in some relevant respect. Thus, to say that a hypothesis or is approximately true is to say that the hypothesis is sufficiently similar to the true hypothesis.

This idea can be formalized if we suppose that there is a (context-appropriate[8]) verisimilitude measure, v, that takes as its input a hypothesis θ and has as its output some real number that represents how similar θ is to the truth. If we presume that such functions are available, we can say that θ is approximately true just in case $v(\theta) < \epsilon$, for some suitably chosen ϵ. There are certain requirements that the verisimilitude measure arguably ought to obey. For example, it arguably ought to be non-negative, and it is also natural to demand that it be continuous whenever the hypothesis space is indexed by a real-valued parameter.

As a concrete example, one non-negative and continuous divergence measure that has been suggested as a verisimilitude measure in a statistical context is the Kullback-Leibler divergence (Forster and Sober, 1994). Supposing that q is the "true" probability distribution that governs the distribution of the data, then the verisimilitude (according to the K-L divergence) of some hypothesis θ (that does not contain adjustable parameters) is $KL(\theta) = -\int q(x) \log \frac{q(x)}{p(x|\theta)} dx$.

Unfortunately, the various ways one might try to accommodate approximate truth within the Bayesian framework face a severe difficulty having to do with the third probability axiom. Briefly, the problem is that, given a set of hypotheses indexed by a parameter, there will generally be multiple parameter values that meet any verisimilitude threshold we set for "approximate truth." Hence, the different parameter values will not be mutually incompatible in the sense that it will be possible for several of them to be approximately true simultaneously. However, Bayesian infer-

[7] See Niiniluoto (1998) for a survey.

[8] In general I agree with Northcott (2013) that there is little reason to assume a priori that there will be a single distance measure that appropriately measures approximate truth in all contexts.

ence requires that the different parameter values be mutually incompatible. Thus, Bayesian inference will in general be impossible if we change the goal of inference from truth to approximate truth. For a more thorough discussion of these issues, and how exactly a conflict with the third probability axiom is to blame, see the appendix.

The underlying problem is that approximate truth is too coarse-grained a concept since it fails to distinguish between several hypotheses, all of which are approximately true. This problem should motivate us to look for an alternative solution to the interpretive problem.

6 The Verisimilitude Interpretation of Probability

Presumably some hypotheses that are approximately true are closer to the truth than other ones, and – at least in many cases – one of the hypotheses under consideration will be closer to the truth than all the others. This suggests a different interpretation of probability. In particular, it is tempting to interpret $p(\theta)$ as the probability that θ is *closest to the truth* out of the hypotheses in Θ; note that in contrast to both truth and approximate truth, closeness to the truth is fundamentally a comparative notion. I will call this interpretation the "verisimilitude interpretation" of probability, and I will use p_c with a c subscript whenever this is the intended interpretation. It is helpful to write out all of the probability axioms with the new interpretation made explicit:

> 1C. $p_c(\Theta) = 1$. Interpretation: one of the hypotheses in Θ is closest to the truth.
>
> 2C. $p_c(\theta) \geq 0$ for all θ. Interpretation: no hypothesis has a negative probability of being closest to the truth.
>
> 3C. $p_c(\bigvee \theta_i) = \sum p_c(\theta_i)$, whenever it is impossible for more than one θ_i to be closest to the truth.

There are several things to note here. First, and most importantly, just about *any* set of hypotheses will satisfy the verisimilitude interpretation of the probability

axioms. More precisely, given any set of hypotheses that can be compared using some verisimilitude measure, at least one of the hypotheses must be maximally close to the truth according to the verisimilitude measure, so the set of hypotheses will satisfy 1C. Hence, the verisimilitude interpretation avoids the interpretive problem of the standard interpretation, which we saw was really a problem with the first axiom.

The verisimilitude interpretation also avoids the problems with the third probability axiom that we identified with the approximate truth approach. In order for Bayesian inference to be possible on the set of hypotheses, the hypotheses must be mutually incompatible in the sense of 3C; that is, it must be impossible for more than one of the hypotheses to be closest to the truth. This axiom will not always be satisfied. For example, if the hypotheses are models and some of the models are contained in others, it may be possible for several of the models to be equally close to the truth, depending on the verisimilitude measure. However, most of the hypothesis sets that Bayesian statisticians study will satisfy 3C.

Another important thing to note is that, under the verisimilitude interpretation, the probability of a hypothesis is always relative to the set of competing hypotheses under consideration. For example, in the set $\{H_1, H_2\}$, $p_c(H_1)$ is the probability that H_1 is closer to the truth than H_2. On the other hand, in the set $\{H_1, H_3\}$, $p_c(H_1)$ is the probability that H_1 is closer to the truth than H_3. The probability of H_1 is, of course, also relative to the verisimilitude measure. The verisimilitude probability of a hypothesis is therefore not an absolute number; it is context-dependent and contrastive. This is in sharp contrast to the standard Bayesian probability of a hypothesis.

Finally, note that $p_c(\theta)$ describes an epistemic attitude different from a degree of belief in the truth of some proposition. Some might be tempted to interpret $p_c(\theta)$ as a standard probability that attaches to the proposition $<\theta$ is closest to the truth$>$. However, this is a mistake, for the reasons mentioned earlier. The proposition $<\theta$ is closest to the truth$>$ belongs to a different algebra than θ does. θ indexes a set of hypotheses in a statistical model, but $<\theta$ is closest to the truth$>$ does not. If Bayesian inference is to be used on the statistical model that is indexed by θ, the probabilities must be assigned to the parameter θ, not to the associated propositions

$<\theta$ is closest to the truth$>$. Hence $p_c(\theta)$ represents an epistemic attitude towards θ, namely the attitude that θ is closest to the truth out of the hypotheses in Θ.

7 The Verisimilitude Interpretation of Probability is Useful

The verisimilitude interpretation of probability is a logically viable solution to the interpretive problem in the sense that it does not face immediate problems with any of the probability axioms. However, some characteristics of the verisimilitude interpretation may seem objectionable. In particular, the fact that the verisimilitude interpretation makes probability assessments contrastive may be regarded as a serious drawback. Perhaps the appropriate response to the interpretive problem is not to adopt the verisimilitude interpretation, but rather to not use Bayesian methods whenever the hypotheses under consideration are all known to be false. On the other hand, maybe there is an alternative solution to the interpretive problem that is better than the verisimilitude interpretation. In this section and the next, I consider both these alternative responses to the interpretive problem.

In order to determine whether the verisimilitude interpretation is defensible, it is helpful to step back for a moment and ask a more fundamental question: why use Bayesian methods at all? If the benefits of Bayesian methods remain even when he standard interpretation of the probability axioms is replaced with the verisimilitude interpretation, then the verisimilitude interpretation is not just logically viable, but potentially *useful*. The goal of the next subsections is to give a preliminary argument for the claim that the verisimilitude interpretation is useful.

7.1 Why be a Bayesian?

What is the benefit of using Bayesian rather than other statistical methods? Perhaps the greatest selling point of Bayesianism is that the prior distribution gives researchers a principled way of incorporating background information. For example, suppose you are estimating the mass of a small cup of water, and suppose you

model the outcome of your measurement as a likelihood function $p(x|m)$, where x is the outcome of your measurement and m is a possible value of the cup's mass. A standard classical ("frequentist") method of estimating the mass of the cup is to choose as your estimate the value of m that maximizes the probability of the observed measurement. This estimation method is known as "maximum likelihood" estimation.

From a Bayesian point of view, maximum likelihood estimation is essentially equivalent to Bayesian inference with a flat (improper) prior probability function that assigns a non-zero and equal probability density to every possible value of m from $-\infty$ to $+\infty$, because the maximum likelihood estimate will be equal to the estimate that has the highest posterior probability if and only if the prior is flat. Clearly, the prior implicitly used in maximum likelihood estimation neglects to incorporate common sense background information that we have about m, and is therefore – from a Bayesian and intuitive point of view – deficient. For example, the mass of an object cannot be a negative number, so no prior should assign any probability mass to negative values of m. Furthermore, we can be absolutely certain that a small cup of water is not going to weigh more than, say, 1kg, so we can also assign a probability of 0 to all values of m greater than 1kg. Thus, as a minimal requirement, any prior probability distribution we use should be restricted to the interval $[0, 1]$. Of course, we have additional common sense knowledge that allows us to restrict the class of sensible prior distributions further.

The above example shows how even very obvious background information can be incorporated in a Bayesian prior in order to improve the inference. Indeed, at least to Bayesian statisticians and scientists who make use of Bayesian methods, this is probably the single biggest advantage that Bayesianism has over its competitors. But how are you supposed to take into account your background information when you are trying to come up with a prior probability distribution over a class of false hypotheses? Do the advantages of Bayesianism carry over when the goal of inquiry changes from finding the truth to finding the hypothesis that is closest to the truth? In the next subsection, I will suggest that the answer is "yes." Scientists often have background knowledge that they can use to discriminate between false hypotheses

in a principled way. And a good way of incorporating this background knowledge is through the construction of a Bayesian prior.

7.2 Verisimilitude and Background Knowledge

Consider again the example concerning the relationship between barometric pressure and the expected amount of rainfall. Suppose one of the things you know about the relationship between barometric pressure and precipitation is that the expected amount of precipitation is not *very* sensitive to changes in barometric pressure. Throughout the whole possible range of barometric pressure, a small change in barometric pressure will not lead to a drastic change in the amount of expected precipitation.

So far, this is background knowledge about the actual, unknown function relating barometric pressure and precipitation. What consequences does this background knowledge have for inferences about the hypothesis set actually under consideration? Suppose, as before, that the hypothesis set you are considering is the set of linear functions. That is, you model the relationship between precipitation and barometric pressure by the set of linear functions $l(Y) = \alpha X + \epsilon$, where ϵ is a normally distributed error term. Can you use your background knowledge to discriminate between the various false linear hypotheses in a principled way? Arguably, you can. Intuitively, by any reasonable measure of verisimilitude, linear functions according to which expected precipitation is not very sensitively dependent on barometric pressure are going to be closer to the truth than are linear functions that model expected precipitation as very sensitively dependent on barometric pressure.

How can all of this be captured reasonably in a prior probability distribution? Let us first see how you can formally capture your background information. Suppose f is the true (and unknown) functional relationship between precipitation and barometric pressure. Then the background information that precipitation does not depend sensitively on changes in barometric pressure can be modeled as a claim about the partial derivative of f (with respect to the barometric pressure variable). The simplest and least sophisticated way of translating your background information into a

quantitative restriction on f' is to to suppose that f' is bounded by some interval (a, b). Next, the intuition that insensitive linear hypotheses are closer to the truth than sensitive linear hypotheses can be formalized as follows: there is some suitably large interval (a', b') that contains (a, b) such that every linear hypothesis l for which l' is bounded by (a', b') is closer to the truth than every linear hypothesis that does not satisfy this requirement. Now, since $l' = \alpha$, the requirement that l' be bounded by (a', b') reduces to the simple requirement that every $\alpha \in (a', b')$ is closer to the truth than every $\alpha \notin (a', b')$. This, in turn, translates to a simple rational requirement on the prior distribution over α, namely that every $\alpha \notin (a', b')$ be assigned a prior probability of 0.

There are more refined ways of formalizing the background information that expected precipitation does not depend very sensitively on barometric pressure. In particular, if we assume a specific verisimilitude measure, then we can get tighter constraints on α.[9] Furthermore, if the hypotheses under consideration are more complicated (i.e. contain more parameters), then the background information will not lead to rational requirements on the prior distribution as neatly. My goal in this section is not, however, to demonstrate in full generality how to best translate background information into reasonable requirements on prior distributions over false hypotheses. My goal is rather to show that it is possible to do so, and that it is plausibly useful. I defer a more thorough treatment of these issues to another time.

[9]For example, suppose we use the following reasonable albeit crude distance measure as our measure of verisimilitude: if f is the true function over the range (m, n) and l is a linear function, then the verisimilitude of l is $v(l) = Max_{x \in (m,n)} |f(x) - l(x)|$. In this case, if we assume that we know that f is bounded by (a, b), then it is possible to prove that every linear function l whose derivative is bounded by (a, b) is closer to the truth than every linear function whose derivative is not bounded in this way, where closeness to the truth is measured using v. For the sake of space, I omit the proof.

8 The Counterfactual Interpretation of Probability

The preceding section shows that the verisimilitude interpretation of the probability axioms is a potentially useful solution to the interpretive problem. However, it may be that there is another solution to the interpretative problem that is better. Earlier, we examined two candidate solutions to the interpretive problem and found them wanting. However, in a very recent paper, Jan Sprenger (2016) proposes a new and different solution to the interpretive problem that is more promising. Sprenger's solution also involves reinterpreting the probability axioms, but he offers a reinterpretation that is interestingly different from the verisimilitude interpretation. However, as we will soon see, given certain plausible assumptions, the verisimilitude solution and Sprenger's solution are formally inter-translatable.

Sprenger's suggestion is that the probability of a false hypothesis can sensibly be interpreted as a *counter-factual* degree of belief. More precisely, suppose α is a parameter that indexes a set of hypotheses, all of which are known to be false. Then any probability assigned to some particular α_0 should be construed as a degree of belief in α_0 that is *conditional* on the (false) supposition that one of the hypotheses indexed by α is true. In other words, the probability of α_0 is really the *conditional* probability $p(\alpha_0 | \vee \alpha)$, where the condition $\vee\alpha$ is the false disjunction that says that one of the α's is true.

This idea is less abstract than it may seem at first blush. As an illustration, suppose I have a coin in a locked cabinet. The probability that the coin would land heads given that I *were* to toss the coin is 0.5, even if it is false that I ever toss the coin. Similarly, according to Sprenger, we can evaluate the probability that a hypothesis is true given that the false supposition that the world *were* such that one of the hypotheses under consideration is true.

According to Sprenger, the counterfactual interpretation of probability offers a simple solution to the interpretive problem that avoids the "muddy waters of verisimilitude." However, in order to actually evaluate counterfactual probabilities in a principled manner, it seems we have to enter waters that are at least as muddy

as the verisimilitude waters. Consider again the example concerning the set of linear hypotheses relating X (barometric pressure) to Y (precipitation in the next hour). We have already agreed that your actual degrees of belief in all of these linear hypotheses is 0. Your degree of belief (or probability density, rather) in some particular linear hypothesis conditional on the disjunction of all the linear hypotheses may still be different from 0, but how are you supposed to figure out what it is? You somehow have to figure out what your probabilities *would* be on the assumption that the world were such that barometric pressure and precipitation were perfectly linearly related. In order for the counterfactual interpretation of probability to be a viable alternative, guidance on how to evaluate counterfactual probabilities is necessary, in the same way that some assumptions about verisimilitude are necessary in order for the verisimilitude interpretation to be viable.

The standard way of evaluating ordinary counterfactuals is by appealing to possible worlds. According to (a simplified version of) Lewis's analysis of counterfactuals (Lewis, 1973), in order to evaluate a counterfactual such as "If A were the case, then B would be the case," you have to go to the closest possible world in which A is true, and then see whether B is true in that world. Crucially, Lewis's analysis depends on a ranking of worlds, where worlds are ranked by how similar they are to the actual world.

Presumably counterfactual probabilities should be assessed in a similar manner. It is not hard to imagine very strange and fanciful possible worlds in which barometric pressure and precipitation are linearly related, but presumably most of those possible worlds are not interesting or relevant. As is the case in counterfactual analysis of conditionals, it is presumably the closest possible worlds that are the interesting ones. But which possible worlds are those? To answer this question, you need to be able to rank worlds in terms of their closeness or similarity to the actual world. But a ranking of possible worlds is hardly easier to come up with than a verisimilitude ranking of hypotheses.

8.1 Relationship Between the Verisimilitude and Counterfactual Solutions

Indeed, in general, any similarity ranking on possible worlds straightforwardly induces a natural verisimilitude ranking on hypotheses, and vice versa.[10] More precisely, suppose we are given a similarity ranking on worlds $w_\alpha \geq w_1 \geq w_2 \geq \ldots$, where w_α is the actual world. Then we can define a verisimilitude ranking on hypotheses as follows: suppose w is the closest world in which H is true and w' is the closest world in which H' is true, then $v(H) \geq v(H')$ if and only if $w \geq w'$.[11]

Conversely, any verisimilitude ranking induces an ordering of possible worlds. Suppose $v(H_0) > v(H_1) > v(H_2) > \ldots$ is a verisimilitude ranking of hypotheses, and for any hypothesis p, let S_p denote the set of worlds in which p is true. Then we can define an ordering of possible worlds in the following way: suppose H is the hypothesis with the highest verisimilitude such that that $w \in S_H$ and suppose H' is the hypothesis with the highest verisimilitude such that $w' \in S'_H$, then $w \geq w'$ if and only if $v(H) \geq v(H)'$.

Thus, although they appear very different, the verisimilitude interpretation and the counterfactual interpretation of probability are formally inter-translatable.

Although the two approaches are formally inter-translatable, they provide different perspectives and help illuminate each other. In particular, it is arguably easier to come up with a verisimilitude measure than a ranking over possible worlds; for example, the Kullback-Leibler measure is a well known verisimilitude measure over statistical models, and this verisimilitude measure will induce a partial ranking over possible worlds. Thus, the verisimilitude approach helps explain where rankings over possible worlds are supposed to come from.

On the other hand, the counterfactual approach helps explain several features of the verisimilitude interpretation as well. For example, earlier we saw that the

[10] For simplicity, the following informal demonstration presupposes the so-called "Uniqueness Assumption" according to which, for every A, there is a unique closest possible world in which A is true. This is a strong and implausible assumption. However, the demonstration does not depend on this assumption.

[11] Hilpinen (1976) uses a similar approach to define a specific verisimilitude measure.

verisimilitude probability of a hypothesis H is relative to the set of hypotheses under consideration. If H is considered as part of the set $\{H, H'\}$, the verisimilitude probability of H is the probability that H is closer to the truth than is H'. But if H is considered as part of the set $\{H, H''\}$, the verisimilitude probability of H is the probability that H is closer to the truth than is H'. The counterfactual interpretation clarifies what is going on here. In the first case, the counterfactual probability that corresponds to $p_c(H)$ is $p(H|H \vee H')$; in the second case, the counterfactual probability that corresponds to $p_c(H)$ is instead $p(H|H \vee H'')$. As can be seen, the two counterfactual probabilities are conditional on different disjunctions, and it is therefore not mysterious that the corresponding verisimilitude probabilities are also different.

9 Summary and Future Research

I have argued that the interpretive problem is a serious problem, but that the problem does not necessarily arise just because the statistical model under consideration is wrong; rather, the interpretive problem arises whenever the *hypotheses of interest* are false. Next, focusing on parameter inference, I have argued that the verisimilitude reinterpretation of the probability axioms provides a logically viable and potentially useful solution to the interpretive problem. Finally, I have contrasted the verisimilitude reinterpretation with another reinterpretation due to Jan Sprenger, and I have argued that the two reinterpretations are formally inter-translatable, but that they nevertheless shed interestingly different lights on the interpretive problem and on each other.

Several important questions remain unanswered, however. In particular, I have not discussed the problem of Bayesian model inference or model selection when all the models are all false. Nor have I discussed in any detail how researchers can come up with principled prior probabilities that discriminate between false hypotheses. Finally, I have not said anything about what consequences reinterpreting the probability axioms has for evidential principles like the Likelihood Principle or the Law of Likelihood. All of this is work for the future.

References

Box, George E. P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness." *Journal of the Royal Statistical Society. Series A (General)*, 143, 383–430.

Carnap, Rudolph (1950), *Logical Foundations of Probability*. The University of Chicago Press.

Easwaran, Kenny (2014), "Regularity and Hyperreal Credences." *Philosophical Review*, 123, 1–41.

Festa, Roberto (1999), "Bayesian Confirmation." In *Experience, Reality, and Scientific Explanation* (Maria Carla Galavotti and Alessandro Pagnini, eds.), volume 61 of *The Western Ontario Series in Philosophy of Science*, 55–87, Springer Netherlands.

Forster, Malcolm and Elliott Sober (1994), "How To Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions." *The British Journal for the Philosophy of Science*, 45, 1–35.

Gelman, Andrew and Cosma Rohilla Shalizi (2013), "Philosophy and the Practice of Bayesian Statistics." *British Journal of Mathematical and Statistical Psychology*, 66, 8–38.

Hilpinen, Risto (1976), "Approximate Truth and Truthlikeness." In *Formal Methods in the Methodology of Empirical Sciences*, number 103 in Synthese Library, 19–42, Springer Netherlands.

Lewis, David K. (1973), *Counterfactuals*. Blackwell Publishers.

Morey, Richard D., Jan-Willem Romeijn, and Jeffrey N. Rouder (2013), "The Humble Bayesian: Model Checking From a Fully Bayesian Perspective." *British Journal of Mathematical and Statistical Psychology*, 66, 68–75.

Niiniluoto, Iikka (1986), "Truthlikeness and Bayesian Estimation." *Synthese*, 67, 321–346.

Niiniluoto, Iikka (1998), "Verisimilitude: The Third Period." *British Journal for the Philosophy of Science*, 49, 1–29.

Northcott, Robert (2013), "Verisimilitude: A Causal Approach." *Synthese*, 190, 1471–1488.

Popper, Karl (1963), *Conjectures and Refutations: The Growth of Scientific Knowledge*. London, Hutchinson.

Pruss, Alexander R. (2014), "Infinitesimals Are Too Small for Countably Infinite Fair Lotteries." *Synthese*, 191, 1051–1057.

Shaffer, Michael J. (2001), "Bayesian Confirmation of Theories That Incorporate Idealizations." *Philosophy of Science*, 68, 36–52.

Sober, Elliott (2015), *Ockham's Razors: A User's Manual*. Cambridge University Press.

Sprenger, Jan (2016), "Conditional Degree of Belief." Manuscript.

Wenmackers, Sylvia and Leon Horsten (2013), "Fair Infinite Lotteries." *Synthese*, 190, 37–61.

A Approximate Truth and Bayesianism

There are two natural ways of trying to accommodate approximate truth within Bayesianism. The first way is to expand the algebra of propositions that p ranges over, so that it also ranges over propositions such as $<\theta$ is approximately true$>$ – or P_θ for short. Thus, even though strictly speaking we assign each θ a probability of 0 of being true, we can consistently assign its associated proposition P_θ a non-zero probability, and moreover this probability represents the probability that P_θ is true

and not just approximately true, since the approximation claim is in the proposition itself. This way, the standard Bayesian interpretation of the probability axioms is preserved.

The other natural way of attempting an accommodation is to abandon the standard Bayesian interpretation of the probability axioms, so that $p(\theta)$ is interpreted as the probability that θ is approximately true rather than true. This line of reasoning is pursued by Niiniluoto (1986) and Festa (1999). Let p_a be a potential probability function where the a subscript indicates that the intended interpretation of $p_a(\theta)$ is the probability that θ is approximately true rather than true. For concreteness, we may imagine that p_a represents the degrees of belief that some agent has in all hypotheses (and models, theories, etc) that the agent takes to potentially be approximately true. By contrast, p can be taken to represent the same agent's degrees of belief in propositions that the agent takes to potentially be true.[12] p_a is therefore defined over a much more expansive set of hypotheses, models, theories, etc. than is p. However, if we allow propositions such as $<\theta$ is approximately true$>$, then presumably there will be a simple correspondence between p_a and p in that we should have $p_a(\theta) = p(P_\theta)$.

There is some reason to prefer working with p_a rather than with propositions such as P_θ. Bayes's formula requires that we assign unconditional probabilities to data x. If we stay inside the original distribution p, this means we have to calculate $p(x) = \sum p(x|P_{\theta_i})p(P_{\theta_i})$, but then we are faced with having to make sense of $p(x|P_{\theta_i})$, or in other words the probability of x conditional on the assumption that θ_i is approximately true. But this is hard to make sense of. In statistical practice, each θ_i will, as was mentioned earlier, in general be part of a fully specified statistical model, which means it will entail a probability for each of the possible outcomes. The associated proposition, P_{θ_i}, however, does not entail any probabilities for data, and it is hard to see how to come up with reasonable conditional probabilities of the form $p(x|P_{\theta_i})$. One might try to argue that it is reasonable to hold that $p(x|P_{\theta_i}) \approx p(x|\theta)$, and this will provide a rough value for $p(x|P_{\theta_i})$, but not a precise one.

[12] Although I hasten to add that a subjective Bayesian perspective will not really play any significant role here.

If, on the other hand, we move to the distribution p_a, then we can expand the probability of x as $p_a(x) = \sum p_a(x|\theta_i)p_a(\theta_i)$. Now, if we suppose that the statistical model stays the same, then it is reasonable to suppose that $p_a(x|\theta_i) = p(x|\theta_i)$; i.e. θ_i still entails the same probability for x in the p_a distribution as it does in the p distribution. The final thing we need to do is to define the joint probability of θ_i and x, which we can naturally define as follows: $p_a(\theta_i \& x) = p(x|\theta_i)p_a(\theta_i)$. Thus, we can write $p_a(x) = \sum p(x|\theta_i)p_a(\theta_i)$.

Introducing the p_a distribution has problems of its own, however, since it's not immediately clear whether such a function can actually satisfy the probability axioms. For example, physicists use both the liquid drop model (L) and the shell model (S) of the nucleus in order to generate predictions, even though these models are logically inconsistent. Presumably, both L and S should be taken to be "approximately true" since they are both auxiliary assumptions used by scientists to generate predictions; hence we should expect it to be the case (at least) that $p_a(L) > 0.5$ and $p_a(S) > 0.5$. However, since L and S are logically inconsistent, the third axiom tells us that $p_a(S \vee L) = p_a(S) + p_a(L) > 0.5 + 0.5 = 1$, which is impossible because (by the first axiom) no probability can be greater than 1. Thus, there is apparently a very foundational problem with trying to change our interpretation of probability so that probabilities are interpreted as probabilities of approximate truth rather than probabilities of truth.

However, on closer inspection, this objection fails. The third probability axiom applies to sets of "logically incompatible" hypotheses; but what does it mean for a set of hypotheses to be logically incompatible? On the standard interpretation, it means that it is not possible for more than one of the hypotheses to be true; i.e. the third axiom is interpreted as follows:

> 3S. $P(\theta_i) = \sum P(\theta_i)$ whenever it is impossible for more than one θ_i to be true.

However, in contexts where approximate truth rather than strict truth is the target, this is arguably not how the axiom should be interpreted. Instead, the axiom should be interpreted in the following way:

> 3A. $P_a(\theta_i) = \sum P_a(\theta_i)$ whenever it is impossible for more than one θ_i to be approximately true.

On the new reading, the earlier objection loses its grip, for – as was pointed out earlier – it is possible for both the shell model and the drop model to be approximately true, so the condition for applying the formula in the third axiom is not met—the two models are not logically incompatible in the sense of $3A$.

Unfortunately, this feature also leads to a serious problem, because the hypothesis spaces that scientists generally use will not be logically incompatible in the sense of axiom $3A$, precisely because it will in general be possible for multiple hypotheses in the hypothesis space to be approximately true. But this is bad news, because in order for Bayes's formula to be applicable, the hypothesis space we use *must* consist of logically incompatible hypotheses, since the denominator of Bayes's formula requires that $p_a(x)$ (or $p(x)$) be expanded in terms of hypotheses that are logically incompatible. Consider, for concreteness, the class of one-variable linear hypotheses, $y = ax$, indexed by the parameter $a \in \mathbb{R}$, and suppose we have available a continuous verisimilitude measure v. Now suppose the true relationship between y and x is not actually linear. Suppose moreover that we set the approximation threshold at $\epsilon > 0$, so that $y = ax$ counts as approximately true if and only if $0 < v(a) < \epsilon$, i.e. if and only if $v(a)$ is in the open interval $S = (0, \epsilon)$. Then the set of hypotheses that are approximately true is indexed by $A = \{a \in \mathbb{R} \mid v(a) \in S\}$. Moreover, $v^{-1}(S) = \{a \in \mathbb{R} \mid v(a) \in S\} = A$, which means A is also an open interval because v is continuous. Since A is an open interval, it has either no members or infinitely many. But this means either none or infinitely many of the hypotheses will be approximately true. In neither case will Bayesian inference be possible. If *none* of the hypotheses are approximately true, then clearly the goal of the inference cannot be to find a hypothesis that is approximately true. If, on the other hand, infinitely many of the hypotheses under consideration count as approximately true, then the hypotheses cannot be used to calculate an unconditional probability for x. But from this it follows that Bayes's formula cannot be applied, and so Bayesian inference will not be possible.

The above problem arises whenever the verisimilitude measure v is continuous and the hypotheses we are considering are parameterized by a real-valued parameter. But many of the hypotheses spaces that applied statisticians make use of *are* parameterized by continuous parameters; hence the problem arises very widely.

There are, as far as I can see, two ways we can try to get out of this problem. As was mentioned earlier, there are two ways the unconditional probability of x can be calculated, depending on whether we use p_a or p with an expanded algebra of propositions. In the p_a distribution we have $p(x) = \sum p_a(x|\theta_i)p_a(\theta_i)$. In the p distribution, we instead have $p(x) = \sum p(x|P_{\theta_i})p(P_{\theta_i})$, where P_{θ_i} is the proposition $<\theta_i$ is approximately true$>$.

If we expand the unconditional probability of x in the first way, we can try to coarse-grain the hypothesis space; if we expand the unconditional probability of x in the second way, we can try to create a partition out of the P_{θ_i} propositions. Neither alternative is very promising.

Let us consider the second way out first. Carnap (1950) taught us how to create a partition out of any set of propositions. The method is as follows: given any set of propositions – A and B, let's say – we form the *state descriptions* $A\&B$, $A\&\neg B$, $\neg A\&B$, $\neg A\&\neg B$. The resulting state descriptions then form a partition. Now, given a set of hypotheses $\{\theta_i\}$, Carnap's method can be used to make a partition out of the set of associated propositions, $<\{\theta_i$ is approximately true$\}>$; the resulting state descriptions will then be logically incompatible (in the sense of $3S$), and we can therefore use Bayes's formula on the resulting partition of state descriptions. There are, however, two major problems with this proposed solution. First, note that if there are n hypotheses in the hypotheses set, then the partition of state descriptions will have 2^n propositions. But that means that if the hypothesis space is parameterized by a continuous parameter – so that its cardinality is $\aleph_1$ – the partition of state descriptions will have cardinality $2^{\aleph_1}$. But it is not possible to assign a regular probability (density) distribution over a set with cardinality $2^{\aleph_1}$. The resulting probability distribution will have to make use of "hyperreal" numbers (Wenmackers and Horsten, 2013), but there are significant difficulties associated with hyperreal probabilities—see, e.g., Easwaran (2014) and Pruss (2014).

The other problem is perhaps even worse. In order to do use Bayes's formula, Bayesians who make use of the above proposed solution will have to somehow assign likelihoods to each of the state descriptions, each of which is a heinous conjunction of propositions of the form $<\{\theta_1$ is approximately true$\}>\&<\{\theta_2$ is approximately true$\}>\&\neg<\{\theta_3$ is approximately true$\}>\&\ldots$etc. It is very hard to see how reasonable probabilities can be assigned conditional on such complicated expressions.

The other possible way out of the problem is to coarse-grain the hypothesis space. If the hypothesis space is parameterized by a continuous parameter, then – as we have seen – infinitely many hypotheses will in general count as approximately true if any hypothesis counts as approximately true. However, if we make the hypothesis space *discrete* by throwing out most of the hypotheses, then the remaining hypotheses may well all be logically incompatible (in the sense of $3A$). For example, if the parameter that indexes the hypotheses ranges over the interval $(0, 1)$, then we could coarse-grain the parameter to $(0.2, 0.4, 0.6, 0.8, 1.0)$, which may well range over hypotheses that are logically incompatible. However, coarse-graining the hypothesis space in this way is not very attractive because (1) how to coarse-grain the space would depend on which ϵ threshold we use, (2) there are multiple ways to coarse-grain a hypothesis space, and each way arbitrarily throws out most of the viable hypotheses. Needless to say, no Bayesian statisticians actually coarse-grain the hypothesis spaces they use in this way; nor, for that matter, do they create state descriptions in the way suggested in the previous solution. Hence, accommodating approximate truth within the Bayesian framework does not seem to be feasible when the hypothesis space is indexed by a continuous parameter.

The above considerations do not show that all is lost for the approximate truth interpretation of probability, however. In particular, if the hypothesis space is discrete, then the above problems may not arise. On the other hand, the problems will arise even with discrete hypotheses spaces, provided there are multiple hypotheses that all meet the verisimilitude threshold that is set for approximate truth. So to prevent these problems from arising, it is necessary to make sure that the hypotheses (or models) under consideration are sufficiently distinct from each other so that only (and precisely) one of them will count as approximately true. Otherwise,

Bayesian methods will not be applicable because the hypotheses (or models) will not be mutually exclusive in the requisite sense (i.e. in the sense of 3A).

But this is an awkward problem to have to deal with. And it points to a defect with the concept of approximate truth: approximate truth is intrinsically too coarse-grained a concept since it fails to distinguish between several hypotheses, all of which are approximately true.

⋆

Establishing causal claims in medicine

⋆

Jon Williamson

Draft of October 28, 2016

Abstract

Russo and Williamson (2007) maintain that in order to establish a causal claim in medicine, one normally needs to establish both that the putative cause and putative effect are appropriately correlated and that there is some underlying mechanism that can account for this correlation. I argue that, although this thesis conflicts with the tenets of contemporary evidence-based medicine (EBM), it offers a better causal epistemology than that provided by EBM because it better explains two key aspects of causal discovery. First, it better explains the role of clinical trials in establishing causal claims. Second, it provides a better account of the logic of extrapolation.

§1 An epistemological thesis

Russo and Williamson (2007, §§1–4) put forward an epistemological thesis that can be phrased as follows:

> In order to establish a causal claim in medicine one normally needs to establish two things: first, that the putative cause and effect are appropriately correlated; second, that there is some mechanism which explains instances of the putative effect in terms of the putative cause and which can account for this correlation.

This epistemological thesis, which has become known as the Russo-Williamson thesis or RWT, has generated some controversy—see, e.g., Weber (2007, 2009); Campaner (2011); Clarke (2011); Darby and Williamson (2011); Gillies (2011); Illari (2011); Howick (2011a,b); Russo and Williamson (2011a,b); Campaner and Galavotti (2012); Claveau (2012); Dragulinescu (2012); Clarke et al. (2013, 2014) and Fiorentino and Dammann (2015). The aim of this section is to explain what the thesis says and why it is controversial. In §2, I argue that an approach to medical methodology based on RWT fares better than present-day EBM in explaining three basic facts about how clinical studies can be used to establish causal claims in medicine. In §3, I argue that it gives a better account of extrapolation inferences too.

First, let us clarify some of the terms that occur within the statement of the thesis. 'Mechanism' here can be understood broadly as referring to a complex-systems mechanism, a mechanistic process, or some combination of the two. A

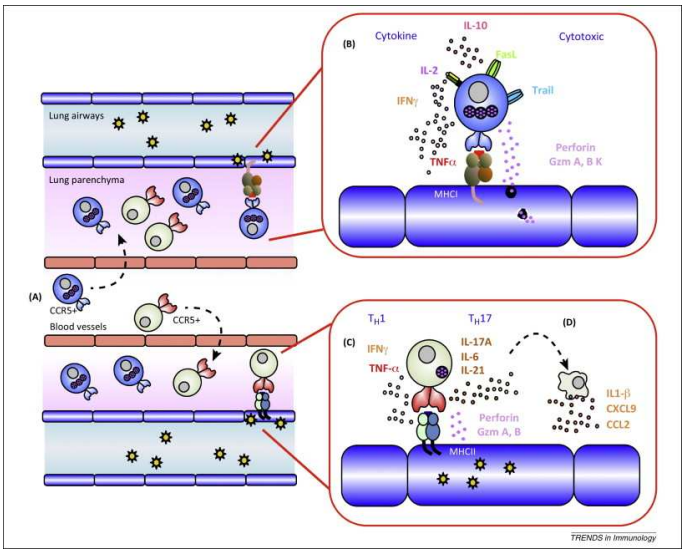


Figure 1: T cell effector mechanisms in a lung infected by influenza A virus (Gruta and Turner, 2014).

complex-systems mechanism consists of entities and activities organised in such a way that they are responsible for some phenomenon to be explained (Machamer et al., 2000; Illari and Williamson, 2012). An example is the mechanism by which the heart pumps blood. A *mechanistic process* is a spatio-temporally contiguous process along which a signal is propagated (Reichenbach, 1956; Salmon, 1998). An example is an artificial pacemaker's electrical signal being transmitted along a lead from the pacemaker itself to the appropriate part of the heart. A mechanism might also be a composed of both these sorts of mechanisms: for example, the complex-systems mechanism of the artificial pacemaker, the complex-systems mechanism by which the heart pumps the blood and the mechanistic process linking the two.

Note that a mechanism is not simply a causal network. A causal network can be represented by a directed graph whose nodes represent events or variables and where there is an arrow from one node to another if the former is a direct cause of the latter. On the other hand, a mechanism is typically represented by a richer diagram, such as that of Fig. 1, in which organisation tends to play a crucial explanatory role. Organisation includes both spatio-temporal structure and the hierarchical structure of the different levels of the mechanism.

Note too that high-quality evidence of mechanism can be obtained by a wide variety of means. Table 1 provides some examples.

Let us turn to some other terms that occur in the epistemological thesis RWT. A causal claim is 'established' just when community-wide standards for granting the causal claim are met. This requires not only high confidence in the truth of the claim itself, but also high confidence in its stability, i.e., that further evidence will not call the claim into question. One theory of evidence holds that evidence consists of propositions that are rationally granted (Williamson, 2015); in which case, when a causal claim is established it can be treated as evidence for other claims.

Table 1: Examples of sources of evidence of mechanisms in medicine (Clarke et al., 2014).

Direct manipulation: e.g., in *vitro* experiments
Direct observation: e.g., biomedical imaging, autopsy, case reports
Clinical studies: e.g., RCTs
Confirmed theory
Analogy: e.g., animal experiments
Simulation: e.g., agent-based models

Establishing is thus crucial to our inferential practice. Establishing requires meeting a high epistemological standard. In particular, establishing a causal claim should be distinguished from acting in accord with a causal claim as a precautionary measure: in certain cases in which a proposed health action has a relatively low cost, or failing to treat has a high cost, it may be appropriate to initiate an action even when its effectiveness has not been established, so that benefits can be reaped in case the effectiveness claim turns out to be true. Despite being a high epistemological standard, establishing—unlike knowing, for instance—is fallible. One can say 'They established that stress is the principal cause of stomach ulcers but further investigations showed they were mistaken,' although one cannot substitute 'knew' for 'established' in this sentence; one would need 'thought they knew' instead.

The epistemological thesis says that one needs to establish that the putative cause and effect are 'appropriately correlated'. Here 'appropriately correlated' just means *probabilistically dependent conditional on potential confounders*, where the probability distribution in question is relative to a specified population or reference class of individuals. Thus, if A is the putative cause variable, B the putative effect variable and C is the set of potential confounder variables, one needs to establish that A and B are probabilistically dependent conditional on C, often written $A \not\perp\!\!\!\perp B \mid C$. A confounder is a variable correlated with both A and B, e.g., a common cause of A and B (Fig. 2). The dependence needs to be established conditional on confounders because otherwise an observed correlation between A and B might be attributable to their correlation with C, rather than attributable to A being a cause of B. The set of *potential* confounders should include any variable that plausibly might be a confounder, given evidence of the area in question. Establishing correlation is non-trivial for two reasons. First, because it requires establishing a probabilistic dependence in the data-generating distribution, rather than simply in the distribution of a sample of observed outcomes. The method of sampling and size of sample can conspire to render an observed sample correlation a poor estimate of a correlation in the population at large. Second, establishing correlation requires considering all potential confounders, and there can be very many of these.

To be clear, we shall use 'observed correlation' to refer to a correlation found in the data, 'genuine correlation' to refer to a correlation in the target population from which the data are drawn, and 'established correlation' to refer to claimed genuine correlation that has met the standards required for being considered established. Since establishing is fallible, that a correlation is established does not deductively imply that there is a genuine correlation, though it makes it very likely. Moreover, to establish a correlation it is not necessary that *every* relevant dataset

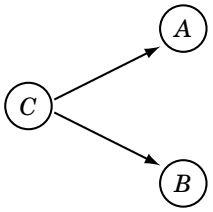


Figure 2: C is a confounder.

yields an observed correlation, although some observed correlation would typically be required.

RWT says that one 'normally' needs to establish both correlation and mechanism. This is because there are certain cases in which causality is not accompanied by a correlation and there are cases in which causality is not accompanied by an underlying mechanism. In cases of overdetermination, where the cause does not raise the probability of the effect because the effect will happen anyway, there is no actual correlation between the cause and the effect. In many such cases, one can expect a *counterfactual* correlation: if things had been different in such a way that the effect would not have happened anyway—e.g., had a second, overdetermining cause been eliminated—then the cause and effect would indeed be correlated. One might think, then, that one ought to be able to establish a counterfactual correlation for any causal claim, if not an actual correlation. However, there are cases in which the cause of interest and a second, overdetermining cause are mutually exclusive, so that it is not possible both to eliminate the second cause and allow the first cause to vary so as to establish a correlation (see Williamson, 2009, §10). So, even the demand for a counterfactual correlation may be too strong. Let us turn next to causality without mechanisms. Where the cause and / or the effect is an absence, it cannot be connected by an actual mechanism. In many such cases, one can expect a *counterfactual* mechanism. Suppose cause and effect are both absences: e.g., failing to treat causes a lack of a heartbeat. If things had been different in such a way that what was absent in the cause were present (e.g., the treatment is administered), then one would expect a mechanism from this presence to a presence corresponding to the effect (e.g., a heartbeat). One might think, then, that one ought to be able to establish the existence of a counterfactual mechanism for any causal claim, if not an actual mechanism. However, there are cases where one of the cause and effect is an absence and the other is a presence, and this strategy does not work. For example, suppose that failing to treat causes a blood clot. That the cause is an absence precludes a mechanism here, but the effect being an absence precludes a mechanism in the obverse case, namely, treating causes an absence of a blood clot. Now, establishing causality in these cases is not particularly problematic in practice. However, it is more subtle than simply establishing correlation and establishing mechanism, even where counterfactual correlations or mechanisms are admitted. The question as to how RWT needs to be modified to cover such cases will be not be considered here, because it is not central to the following arguments. The use of 'normally' is intended to leave open the possibility that in certain tricky cases one might not need to establish both correlation and mechanism.

RWT requires establishing the *existence* of a correlation and the *existence* of a mechanism, not the extent of the correlation, nor the details of the mechanism.

Table 2: Possible explanations of an observed correlation between A and B.

Causation	A is a cause of B.
Reverse causation	B is a cause of A.
Confounding (selection bias)	There is some confounder C that has not been adequately controlled for by the study.
Performance bias	Those in the A-group are identified and treated differently to those in the $\neg A$-group.
Detection bias	B is measured differently in the A-group in comparison to the $\neg A$-group.
Chance	Sheer coincidence, attributable to too small a sample.
Fishing	Measuring so many outcomes that there is likely to be a chance correlation between A and some such B.
Temporal trends	A and B both increase over time for independent reasons. E.g., prevalence of coeliac disease & spread of HIV.
Semantic relationships	Overlapping meaning. E.g., phthiasis, consumption, scrofula (all of which are TB).
Constitutive relationships	One variable is a part or component of the other.
Logical relationships	Measurable variables A and B are logically complex and logically overlapping. E.g., A is $C \wedge D$ and B is $D \vee E$.
Physical laws	E.g., conservation of total energy can induce a correlation between two energy measurements.
Mathematical relationships	E.g., mean and variance variables from the same distribution will often be correlated.

Of course, in some cases establishing the extent of a correlation is a means to establishing its existence, and establishing the details of a mechanism is a means to establishing its existence, but these means are not the only means. We shall return to this point in §2.

RWT is a purely epistemological thesis, concerning the discovery of causal relationships. Russo and Williamson (2007) used the thesis to argue for a particular metaphysical account of causality, the epistemic theory of causality, but RWT itself does not say anything about the nature of causality. The thesis is intended to be both descriptive and normative: i.e., as capturing typical past cases of establishing causality in the biomedical sciences (e.g., Clarke, 2011; Gillies, 2011), as well as characterising the way in which one ought to establish causality.

To see why one ought to establish causality this way, consider that an observed correlation between two variables might be explained in a wide variety of ways, as depicted in Table 2. Some of these explanations provide reason to doubt that there is a genuine correlation in the underlying population. For example, one of the potential confounders might not have been adequately controlled for, or the sample may be rather small. On the other hand, some of these explanations provide reason to doubt that A is a cause of B, even where there is a genuine correlation between

these variables. For example, there might be some variable that could not possibly be considered a potential confounder, given the evidence available, but nevertheless is a confounder, and has not been adequately controlled for. In such a case A and B can be appropriately correlated yet A may not be a cause of B—the correlation is attributable to a common cause. Or there may be a genuine correlation that is entirely non causal, explained by a semantic relationship, for instance. Thus there are two forms of error: error when inferring correlation in the data-generating distribution from an observed correlation and error when inferring that A is a cause of B from an established correlation. Evidence of mechanisms can help to eliminate both forms of error. For instance, it can help to determine the direction of causation, which variables are potential confounders, whether a treatment regime is likely to lead to performance bias, and whether measured variables are likely to exhibit temporal trends.

The existence of the second kind of error—error when inferring that A is a cause of B from an established correlation—shows that it is not enough to simply establish correlation. If it is genuinely the case that A is a cause of B, then there is some combination of mechanisms that explains instances of B by invoking instances of A and which can account for the correlation. Hence, in order to establish efficacy one needs to establish mechanism as well as correlation.

Let us consider an example. The International Agency for Research on Cancer (IARC) Monographs evaluate the carcinogenicity of various substances. When evaluating whether mobile phone use is a cause of cancer, IARC found that the largest study (the INTERPHONE study) showed a correlation between the highest levels of call time and certain cancers. This correlation was confirmed by another large study from Sweden. However, evidence of mechanisms was weak and certainly failed to establish the existence of an underlying mechanism. For this reason, chance or bias were considered to be the most likely explanations of the observed correlations, and while causality was not ruled out, neither was it established (IARC, 2013, §§5–6).[1]

⋆

Further discussion of the descriptive and normative adequacy of RWT can be found in the above references. We will not revisit all these arguments here. Instead, I shall argue here that RWT provides a better account of the epistemology of causality than a rival theory which is motivated by evidence-based medicine (EBM).

Evidence-based medicine is concerned with making the evaluation of evidence explicit:

> Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. (Sackett et al., 1996)

Of course, this goal is hardly controversial. What characterises present-day EBM is not the goal itself but the means by which it attempts to achieve this goal. EBM employs hierarchies of evidence in order to evaluate evidence and these hierarchies of evidence tend to favour clinical studies and statistical analyses of these studies over other forms of evidence. Clinical studies (CSs) measure the putative cause and effect, together with potential confounders. CSs include controlled experiments such as randomised controlled trials (RCTs) as well as observational studies such as

[1] I am very grateful to Julian Reiss for alerting me to this example.

Figure 3: SUNY Downstate Medical Center EBM Tutorial (SUNY, 2004).

cohort studies, case control studies, case series and collections of case reports. In particular, non-CS evidence of mechanisms, i.e., evidence of mechanisms obtained from means other than the clinical studies that seek to establish a correlation between the putative cause and effect, tends to be either ignored or relegated to the bottom of the hierarchy. For example, Fig. 3 depicts an evidence hierarchy of SUNY (2004), used for EBM training. This places animal research and in vitro research, which in the right circumstances can provide high quality evidence of mechanisms, below 'opinions', and well below evidence obtained from clinical studies and statistical analyses of CSs. A similar point can be made about an evidence hierarchy used by the UK National Institute for Health and Care Excellence, depicted in Fig. 4. Fig. 5 depicts the current evidence hierarchy of the Oxford Centre for Evidence-Based Medicine, which places 'mechanism-based reasoning' at the lowest level. Other approaches, such as the GRADE system, tend to overlook evidence of mechanisms entirely (Guyatt et al., 2011, Fig. 2).

The main feature of contemporary EBM of relevance to this paper, then, is that it views non-CS evidence of mechanisms as either irrelevant to the process of evidence evaluation or as strictly inferior to evidence obtained from clinical studies and analyses of CSs. In the latter case, opinions differ as to whether or not CSs trump non-CS evidence of mechanisms, i.e., whether or not one should ignore non-CS evidence of mechanisms when clinical studies are available. Either way, however, the sort of studies that provide high quality evidence of correlation are viewed as superior to other investigations which can provide high quality evidence of mechanism.

As a consequence, contemporary EBM stands in conflict with RWT. EBM prioritises evidence of correlation over other evidence of mechanism, whereas RWT treats evidence of mechanism alongside evidence of correlation. An advocate of RWT might interpret the EBM hierarchy of evidence as a way of evaluating evidence of correlation, rather than causation. Under this interpretation, Fig. 6 portrays the epistemological picture motivated by EBM. On the other hand, Fig. 7 represents the approach motivated by RWT (Clarke et al., 2014).

Given this conflict and the fact that EBM is now very well entrenched in medicine, it is no wonder that RWT is controversial. However, we shall see that

Levels of evidence	Type of evidence
Ia	Systematic review (with homogeneity)[a] of level-1 studies[b]
Ib	Level-1 studies[b]
II	Level-2 studies[c] Systematic reviews of level-2 studies
III	Level-3 studies[d] Systematic reviews of level-3 studies
IV	Consensus, expert committee reports or opinions and/or clinical experience without explicit critical appraisal; or based on physiology, bench research or 'first principles'

[a] Homogeneity means there are no or minor variations in the directions and degrees of results between individual studies that are included in the systematic review.

[b] Level-1 studies are studies:
- that use a blind comparison of the test with a validated reference standard (gold standard)
- in a sample of patients that reflects the population to whom the test would apply.

[c] Level-2 studies are studies that have **only one** of the following:
- narrow population (the sample does not reflect the population to whom the test would apply)
- use a poor reference standard (defined as that where the 'test' is included in the 'reference', or where the 'testing' affects the 'reference')
- the comparison between the test and reference standard is not blind
- case–control studies.

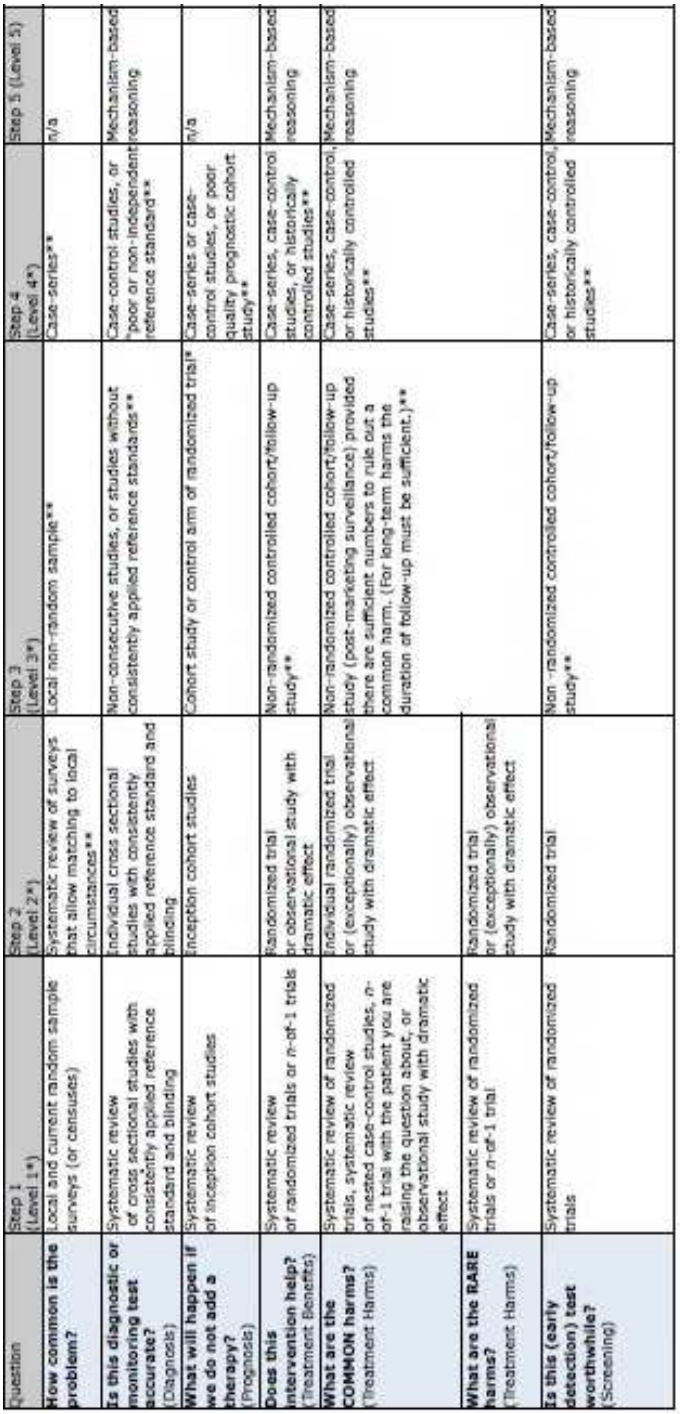
[d] Level-3 studies are studies that have **at least two or three** of the features listed for level-2 studies.

Figure 4: Hierarchy of evidence for diagnostic studies from NICE (2006, p.48).

there are good reasons to prefer the RWT-motivated causal epistemology to the EBM-motivated view. Next, in §2, I shall argue that RWT better explains the role of clinical studies in establishing a causal claim. In §3 I shall argue that RWT better explains the process of extrapolating a causal claim from the study population to a target population.

That present-day EBM fails to provide an adequate epistemology of causality does not imply that the whole enterprise of evidence-based medicine is doomed. Current EBM provides a first approximation to the correct epistemology and has led to numerous advances in patient care. The claim made here is that improvements to contemporary EBM can be made, and that the picture of Fig. 7 provides a better approximation. This picture can thus be viewed as a way to develop 'EBM+', i.e., as a proposal to advance the methodology of EBM by taking better account of evidence of mechanisms (c.f., ebmplus.org). No claim is made that Fig. 7 is the end of the story—further improvements can be made, no doubt.

While present-day EBM advances an essentially monistic account of causal discovery in terms of clinical studies that are evaluated according to how well they establish correlation, the RWT-motivated EBM+ approach is dualistic, treating evidence of mechanisms and evidence of correlation on a par. In this sense, EBM+ has a close affinity to the approaches of Claude Bernard and Austin Bradford Hill, both of whom advocated an approach to medicine which treats evidence of mechanisms



Question	Step 1 (Level 1*)	Step 2 (Level 2*)	Step 3 (Level 3*)	Step 4 (Level 4*)	Step 5 (Level 5)
How common is the problem?	Local and current random sample surveys (or censuses)	Systematic review of surveys that allow matching to local circumstances**	Local non-random sample**	Case-series**	n/a
Is this diagnostic or monitoring test accurate? (Diagnosis)	Systematic review of cross sectional studies with consistently applied reference standard and blinding	Individual cross sectional studies with consistently applied reference standard and blinding	Non-consecutive studies, or studies without consistently applied reference standards**	Case-control studies, or "poor or non-independent reference standard**	Mechanism-based reasoning
What will happen if we do not add a therapy? (Prognosis)	Systematic review of inception cohort studies	Inception cohort studies	Cohort study or control arm of randomized trial*	Case-series or case-control studies, or poor quality prognostic cohort study**	n/a
Does this intervention help? (Treatment Benefits)	Systematic review of randomized trials or *n*-of-1 trials	Randomized trial or observational study with dramatic effect	Non-randomized controlled cohort/follow-up study**	Case-series, case-control studies, or historically controlled studies**	Mechanism-based reasoning
What are the COMMON harms? (Treatment Harms)	Systematic review of randomized trials, systematic review of nested case-control studies, *n*-of-1 trial with the patient you are raising the question about, or observational study with dramatic effect	Individual randomized trial or (exceptionally) observational study with dramatic effect	Non-randomized controlled cohort/follow-up study (post-marketing surveillance) provided there are sufficient numbers to rule out a common harm. (For long-term harms the duration of follow-up must be sufficient.)**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning
What are the RARE harms? (Treatment Harms)	Systematic review of randomized trials or *n*-of-1 trial	Randomized trial or (exceptionally) observational study with dramatic effect			
Is this (early detection) test worthwhile? (Screening)	Systematic review of randomized trials	Randomized trial	Non -randomized controlled cohort/follow-up study**	Case-series, case-control, or historically controlled studies**	Mechanism-based reasoning

Figure 5: Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence (OCEBM Levels of Evidence Working Group, 2011).

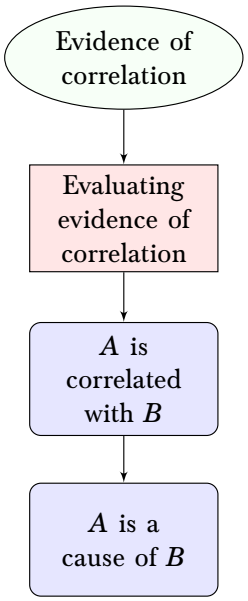


Figure 6: The causal epistemology of contemporary EBM.

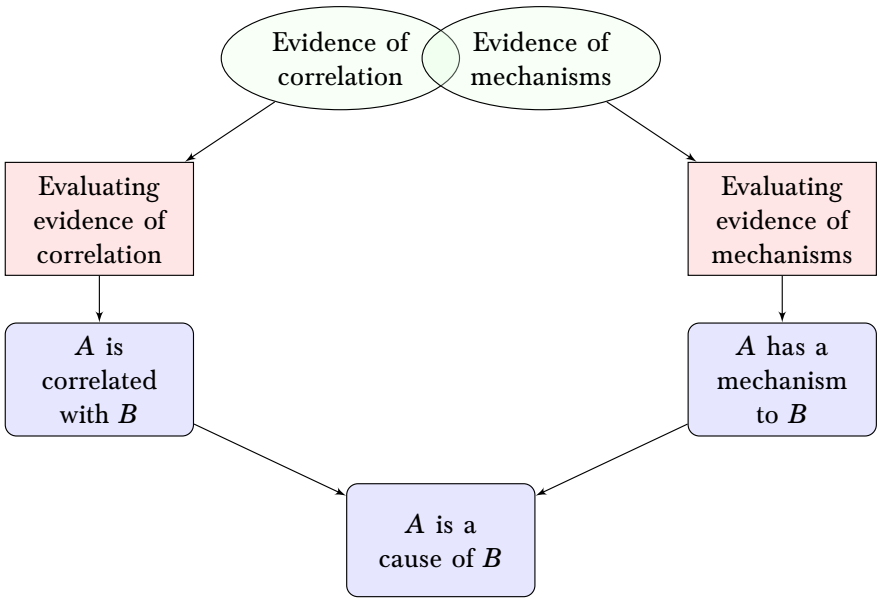


Figure 7: Evidence of mechanisms treated alongside evidence of correlation.

on a par with evidence of correlation (Russo and Williamson, 2007, §2; Clarke et al., 2014, §2.2).

§2 Explananda concerning clinical studies

In this section, I shall argue that RWT can successfully explain three fundamental facts about the role of clinical studies in establishing a causal claim, and that the view motivated by present-day EBM cannot account for all of these facts (although it can account for the first fact). The three facts are these: (i) in some cases, clinical studies suffice to establish a causal claim; (ii) in some cases, randomised studies are not required to establish a causal claim; (iii) in some cases, randomised studies are trumped by other evidence of mechanisms. We shall examine each of these facts in turn.

§2.1. In some cases, clinical studies suffice to establish a causal claim

Howick (2011a) suggests that in a number of cases, medical interventions have been accepted on the basis of comparative clinical studies alone. He cites the following cases: the use of aspirin as an analgesic; the use of general anaesthesia; and the use of deep brain stimulation in treating patients with advanced Parkinson's disease or Tourette's syndrome. He argues that these cases are a problem for the epistemological thesis RWT, because the mechanisms of action were not—in some cases, still are not—known. Howick points out that these cases are quite compatible with contemporary EBM, which focuses overwhelmingly on clinical studies.

In response to this objection, one might question whether, in these examples, the causal claims really were were established on the basis of comparative clinical studies alone. However, I do not want to question the examples here, because I want to accept the general principle that it is possible that clinical studies alone can be used to establish a causal claim in medicine. The point I want to make is that this general principle is quite compatible with RWT.

Consider the RWT-motivated picture of Fig. 7. Some of the total available evidence can be considered to provide evidence of correlation, in the sense that these items of evidence contribute to support or undermine the claim that the putative cause and effect are appropriately correlated. (An item of evidence *contributes* to support a claim if, when taken together with other items, it supports the claim, and the other items do not on their own support the claim to the same degree.) Some of the total available evidence can be considered to provide evidence of mechanisms, in the sense that these items of evidence contribute to support or undermine a claim that there is some mechanism which explains instances of the putative effect in terms of the putative cause and which can account for the extent of the correlation. There is no suggestion that an item of evidence cannot provide both evidence of correlation and evidence of mechanisms.

In particular, clinical studies not only provide evidence of correlation, they can also—in the right circumstances—provide high-quality evidence of mechanisms (Table 1). Suppose:

- There are sufficiently many independent clinical studies,
- They are of sufficient quality (e.g., they are sufficiently large, well-conducted RCTs),

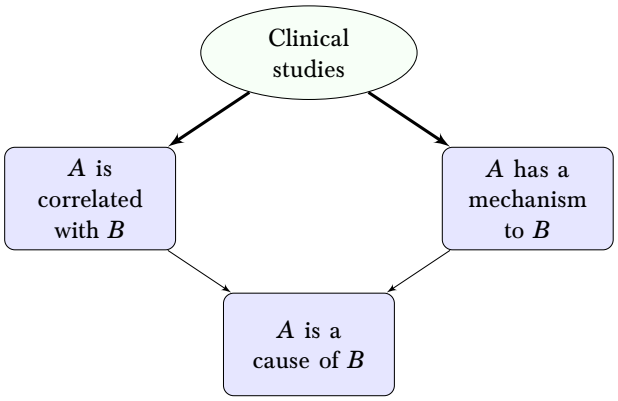


Figure 8: Clinical studies can, in the right circumstances, establish a causal claim.

- Sufficiently many studies point in the same direction,
- They observe a large enough correlation ('effect size'),
- It is clear that the variables in question are such that there are no temporal trends and they are not semantically, constitutively, logically, physically or mathematically related,
- There is no other evidence to suggest lack of a suitable mechanism.

Then one can infer both:

- That there is a correlation,
- That there must be some underlying mechanism that explains this correlation.

This is because alternative explanations of the correlation are ruled out (c.f. Table 2). In such cases, clinical studies can provide evidence of the existence of a mechanism even though they may fail to shed light on the details of the mechanism. Fig. 8 depicts this kind of inference, from the perspective of RWT. (Here an arrow from node X to node Y is thick if X on its own would suffice to establish Y; an arrow is thin if X is insufficient on its own to establish Y, but nevertheless supports Y.) In practice, however, the conditions of this inference are rarely met. Thus, non-CS evidence of mechanisms is typically crucial in establishing causality.

In sum, then, while Howick cites as counterexamples to RWT cases in which clinical studies have sufficed to establish causality, these cases are in fact quite compatible with RWT.

Confusingly, Howick also cites as evidence against RWT a range of cases in which evidence of mechanisms alone led to erroneous causal inferences; see also Howick (2011b, Chapter 10). These cases clearly confirm—rather than disconfirm—RWT, which says that causal claims cannot be established just by establishing mechanisms, since one needs to establish correlation as well. Moreover, these cases also support the EBM+ approach, which holds that evidence of mechanisms needs to made explicit and its quality scrutinised, just as contemporary EBM strives to do with evidence of correlation arising from statistical studies. This is because in many of these cases the evidence of mechanisms was rather weak.

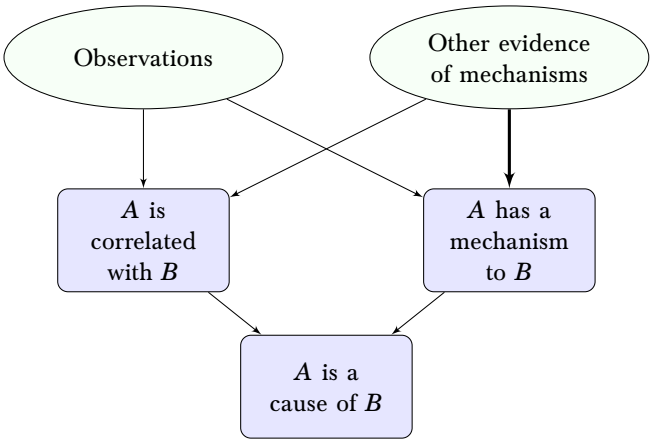


Figure 9: One way to establish a causal claim without RCTs.

§2.2. In some cases, randomised studies are not required to establish a causal claim

To support the view that in some cases there is no need for RCTs when establishing causality, consider three examples.

First consider the tongue-in-cheek conclusions of Smith and Pell (2003), who study 'parachute use to prevent death and major trauma related to gravitational challenge':

> As with many interventions intended to prevent ill health, the effectiveness of parachutes has not been subjected to rigorous evaluation by using randomised controlled trials. Advocates of evidence based medicine have criticised the adoption of interventions evaluated by using only observational data. We think that everyone might benefit if the most radical protagonists of evidence based medicine organised and participated in a double blind, randomised, placebo controlled, crossover trial of the parachute. (Smith and Pell, 2003, p. 1459.)

From the point of view of EBM, the evidence for the effectiveness of parachutes is very weak: no systematic studies, let alone RCTs, and some mechanistic evidence which sits at the bottom of the evidence hierarchy, if it occurs at all. It is hard to see how causality could be established on the basis of this evidence, if EBM is right. From the point of view of EBM+, however, the evidence is strong: excellent evidence of mechanisms, and, although unsystematic, plenty of observational evidence relating to instances where parachutes were and were not used, and a very large observed effect size. From the point of view of EBM+, the evidence of mechanisms on its own suffices to establish the existence of a suitable mechanism, and, when combined with the unsystematic observations, the total evidence suffices to establish correlation too. Hence causality is established. This inference is depicted in Fig. 9.

Having clarified the structure of this inference, let us consider a second example, also considered by Worrall (2007). The question here is how to establish

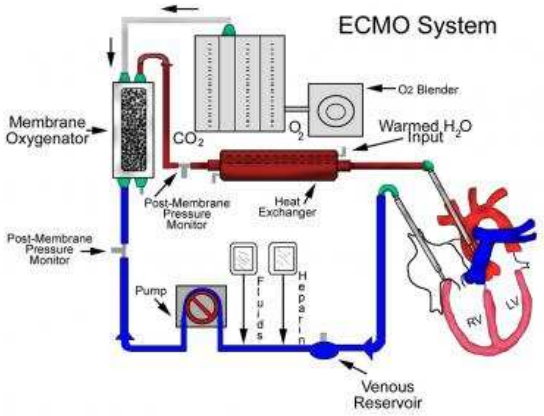


Figure 10: The ECMO mechanism.

the effectiveness of extracorporeal membraneous oxygenation (ECMO) for treating persistent pulmonary hypertension (PPHS). With PPHS, immaturity of the lungs in certain newborn babies leads to poor oxygenation of the blood. ECMO oxygenates the blood outside the body (Fig. 10). Observational studies suggested that ECMO increases survival rate from about 20% to about 80% (Bartlett et al., 1982). However, under standard EBM procedures for evaluating evidence, the available evidence was viewed as insufficient to establish causality, and it was felt necessary to conduct an RCT (Bartlett et al., 1985). At least five subsequent RCTs were carried out, leading to significant loss of life in the control groups.

Conducting RCTs in such a case is considered standard EBM procedure. That non-RCT evidence is viewed as insufficient by EBM was confirmed by a recent Cochrane Review of ECMO, which explicitly disregarded any evidence that did not take the form of an RCT (Mugford et al., 2010).

On the other hand, Worrall (2007) suggests that RCTs were unnecessary in the ECMO case. This conclusion is supported by the RWT-motivated EBM+ approach. This case is analogous to the parachute case: there was strong observational evidence which indicated a large effect size, as well as excellent evidence of mechanisms. Indeed, as in the parachute case, the details of the mechanism of action were very well established. Thus Fig. 9 captures the pre-RCT inference in the ECMO case. There is little doubt that conducting RCTs led to yet greater surety; however, despite being mandated by EBM, RCTs were arguably unnecessary to establish causality.

As a third example, consider the case of establishing the carcinogenicity of aristolochic acid.[2] When IARC originally investigated aristolochic acid in 2002, it found that, while there was observational evidence that Chinese herbs that contain aristolochic acid cause cancer, there was 'limited' evidence in humans concerning the carcinogenicity of aristolochic acid itself as an active ingredient, so carcinogenicity could not be established (IARC, 2002, pp. 69–128). IARC re-examined the question some years later and found that there was little in the way of further

[2]I am very grateful to Kurt Straif for alerting me to this example.

observational evidence in humans, so the study evidence involving humans was still 'limited'. However, there was much more evidence of the underlying mechanisms available, to the extent that the mechanistic evidence could now be described as 'strong' and causality could be considered established (IARC, 2012, pp. 347–361). The key point here is that the change in evidence that warranted establishing causality was a change in mechanistic evidence.

These three cases thus instantiate the following form of inference. Suppose:

- The mechanisms involved are established,
- Observational studies suggest a sufficiently large effect size,
- Sufficiently many studies point in the same direction,
- It is clear that the mechanisms involved can account for the effect size,
- It is clear that the variables in question are such that there are no temporal trends and they are not semantically, constitutively, logically, physically or mathematically related,
- There is no other evidence to suggest that the correlation is best explained in another way,

Then one can infer both:

- That there is a correlation,
- That there must be some underlying mechanism that explains this correlation.

In these cases, evidence of mechanisms obtained by means other than clinical studies provides evidence of correlation. When taken in conjunction with the observational studies, this can be sufficient to establish a correlation (Fig. 9). Note that the observational studies do not need to be very systematic: this is so in the parachute case, and it may also be true when establishing some adverse drug reactions (Aronson and Hauben, 2006; Hauben and Aronson, 2007).

While this mode of inference clearly fits the EBM+ approach, which is motivated by RWT, it is harder for contemporary EBM to explain, because, as we saw in the ECMO case, much of the practice of present-day EBM seems to demand randomised studies in order to establish causality. To be sure, some deny that randomised trials are required. For example, Glasziou et al. (2007) argue that in cases where there is a large effect size, RCTs may be unnecessary. However, they struggle to explain from within the EBM paradigm how evidence of mechanisms can be treated on a par with observational studies to help establish causality. Instead they evoke the Bradford Hill indicators of causality, and the Bradford Hill approach is much more in line with RWT and EBM+ than with standard EBM (§1).

§2.3. In some cases, randomised studies are trumped by other evidence of mechanisms.

So far, we have seen that while present-day EBM can account for situations in which RCTs are sufficient to establish causality, it is doubtful whether EBM adequately handles cases in which RCTs are unnecessary. As we shall now see, it is clear that EBM cannot capture cases in which randomised studies are trumped by other evidence of mechanisms. This is because evidence of mechanisms obtained by

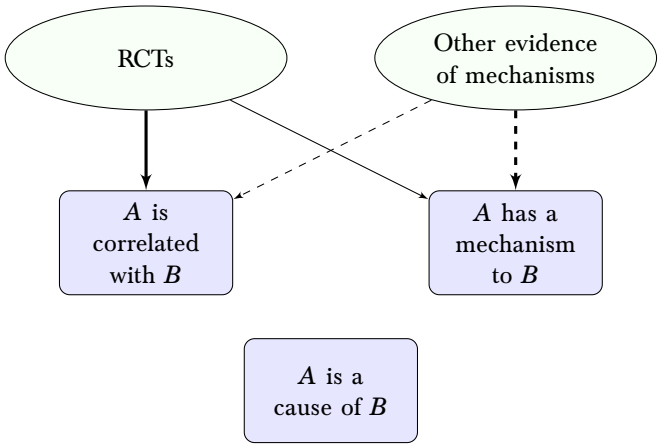


Figure 11: Evidence of a lack of mechanism can trump RCTs.

means other than randomised studies is viewed—when it is considered at all—as strictly inferior to evidence arising from randomised studies (§1).

There are two kinds of example here. One sort of example involves positive evidence of causality from randomised studies; this evidence is trumped by evidence that there is no mechanism by which causality can operate. To start with another tongue-in-cheek example, Leibovici (2001) presented an RCT which observed a correlation between remote, retroactive intercessionary prayer and length of stay of patients in hospital. The patients in question had bloodstream infections in Israel during the period 1990–6; the intervention involved saying 'a short prayer for the well being and full recovery of the group as a whole' in the year 2000 in the USA, long after recovery or otherwise actually took place. The study also found a correlation between the intervention and duration of fever. The author concludes:

> No mechanism known today can account for the effects of remote, retroactive intercessory prayer said for a group of patients with a bloodstream infection. However, the significant results and the flawless design prove that an effect was achieved. (Leibovici, 2001, p. 1451.)

Present-day EBM clearly accords with this inference to an effect, because it views considerations to do with mechanisms as strictly inferior to evidence produced by clinical studies. However, the implicit inference is that this conclusion is ridiculous: no effect should be inferred. This contrary conclusion goes against EBM. It is not possible for present-day EBM to account for the possibility that a large, well-conducted RCT can be trumped by the fact that current science has no place for a mechanism between remote, retroactive intercessionary prayer and length of stay of in hospital. On the other hand, this is quite compatible with EBM+. Fig. 11 depicts the inference here, from the perspective of RWT. Undermining evidence is represented by dashed arrows. The thick dashed arrow depicts an inferential connection that is enough on its own to rule out mechanism. As before, the thick solid arrow depicts a connection that would normally be enough on its own to establish the conclusion (correlation): a significant result from a large well-conducted RCT.

However, there is evidence which undermines this conclusion: well-confirmed scientific theory. The presence of this undermining evidence blocks any inference to either correlation or mechanism, and thereby blocks an inference to causation.

Other inferences follow the same pattern. Some comparative studies for precognition have observed a significant correlation (see, e.g., Bem, 2011), as have others in the case of homeopathy (e.g., Cucherat et al., 2000; Faculty of Homeopathy, 2016). What are the options for resisting an inference to causality in such cases? EBM will point to the fact that the evidence base shows mixed results and is thus inconclusive. However, while this may be so for precognition and homeopathy in general, it is not the case for certain specific interventions which are instances of precognition or homeopathy; as the above references show, there are specific interventions for which only positive studies are available. An alternative way to resist an inference to causality in such cases is to invoke the machinery of Bayesianism: to argue that the prior probability of effectiveness is so low that the posterior probability remains low, despite confirmatory trials. This strategy is open to the charge of subjectivity, or relativity to the way the problem is framed. This charge is clearly correct in the case of a subjective Bayesian analysis, but even under an objective Bayesian reading this charge is on the mark, as an objective Bayesian analysis typically requires high prior probability in deception or experimental error (Jaynes, 2003, §§5.1–2). A third alternative is to apply the RWT-motivated EBM+ approach. The inference in these cases follows the pattern of Fig. 11, and it is clear that causality has not been established, even in specific cases where trials would sufficient in the absence of other evidence to establish correlation. Arguably, then, the RWT-motivated approach is the most promising of these three strategies.

In the kind of example considered above, positive evidence from randomised studies is trumped by evidence of absence of mechanism. But there is another sort of example, in which there is observational evidence, evidence from RCTs and other evidence of mechanisms, and in which the other evidence of mechanisms plays more of a role in establishing causality than do the RCTs. The ECMO case takes this form at the point after the first randomised trial. The first randomised trial provided weak evidence, because after the first baby was randomly assigned to the control arm of the trial and subsequently died, no more individuals were assigned to this arm. Thus the size of the trial was not sufficient to draw any strong conclusions. Arguably, at that point in time the evidence of mechanisms was stronger than that arising from RCTs and it played more of a role in establishing causality. Indeed, if the analysis of §2.2 is correct then the RCT evidence was redundant. The evidence of mechanisms trumps the RCT evidence in such a case.

⋆

To conclude, the causal epistemology motivated by RWT can validate all three facts about the role of clinical studies in establishing a causal claim. The EBM approach certainly captures the first fact (in some cases, clinical studies suffice to establish a causal claim). However, the practice of EBM goes against the second fact (in some cases, randomised studies are not required to establish a causal claim) and EBM certainly fails to explain the third fact (in some cases, randomised studies are trumped by other evidence of mechanisms).

§3
Extrapolation

We now turn to the question of how a causal claim can be extrapolated from a study population to a target population of interest. This mode of inference is ubiquitous, because the population within which clinical studies establish a correlation (e.g., hospital patients in a particular region who are not too young, not too old, not too ill and not pregnant) rarely coincides with the population within which a treatment is intended to be used. It is also very common—and particularly challenging—to extrapolate causal claims from animals to humans. Any adequate causal epistemology needs to explain how extrapolation is possible and needs to clarify the logic of extrapolation.

Here is a first approximation to the logic of extrapolation:

The causal relationship holds in the study population
The study and target populations are similar in causally relevant respects
---
The causal relationship holds in the target population

As Steel (2008) points out, this explication faces two immediate problems. The first, which Steel calls the *extrapolator's circle*, is that 'it needs to be explained how we could know that the model and the target are similar in causally relevant respects without already knowing the causal relationship in the target' (p. 78). The worry is that extrapolation seems redundant, since the conclusion of the above rule of inference is apparently needed to establish the second premiss. The second problem, which we shall call the *extrapolator's block*, is that 'any adequate account of extrapolation in heterogeneous populations must explain how extrapolation can be possible even when [causally relevant differences between the model and the target] are present' (pp. 78–9). That is, the study and target population are rarely entirely similar in causally relevant respects—particularly when extrapolating from animals to humans—and it needs to be made clear what sort of differences are permissible in order to prevent the second premiss of the above argument from failing and the inference thereby being blocked.

Note that the study population is chosen for investigation precisely because it is easier to conduct clinical studies on this population than on the target population. Thus the clinical studies that one can perform on the study population tend to be of a higher standard than those directly obtained on the target population. In the light of this fact, one can sketch an approach to extrapolation motivated by contemporary EBM as follows:

RCTs establish a causal relationship in the study population
Observational studies in the target population are consistent with this relationship
---
The causal relationship holds in the target population

This approach to extrapolation circumvents the aforementioned problems very nicely. There is no extrapolator's circle because one does not need to know that the causal relationship holds in the target population to obtain observational studies in the target population. There is no extrapolator's block because this theory of extrapolation makes extrapolation possible even when there are substantial differences between the study and target populations.

That there may be substantial differences between the study and target populations points to two problems that face the EBM-motivated approach. First we

have what we might call the *extrapolator's fallacy*: it needs to be explained how extrapolation is a reliable form of inference, rather than simply fallacious. The worry is that the EBM-motivated account will lead to lots of mistaken conclusions, because observational studies in the target population typically provide weak evidence that the target population is similar to the study population in causally relevant respects. This problem may explain some recent scepticism about extrapolation amongst those interested in medical methodology (see, e.g., Ioannidis, 2012). However, since almost every causal claim of interest has to be extrapolated from some study population, scepticism is hardly a viable option.

The second, related problem is that the *extrapolator's standards are slipping*. In the EBM-motivated approach we have a high standard for internal validity but a low standard for external validity: evidence deemed to be of high quality by EBM (such as that obtained from RCTs) is used to establish causality in the study population, while low quality evidence (such as that obtained from observational studies) is used to establish causality in the target population. In general, an account of extrapolation should not have double standards—the burden of proof for causality should be similar in the study and target populations.

As Steel (2008, Chapter 5) suggests, in order to extrapolate a causal claim from a study population to a target population, one needs evidence that similar mechanisms operate in the two populations. This is particularly important in contexts where mechanisms are likely to differ, such as with extrapolations from animals to humans or interventions involving long causal pathways. It turns out that this feature of extrapolation can be captured by the following RWT-motivated account.

Fig. 12 depicts an account of the logic of extrapolation that is motivated by RWT. In the study population, one can carry out clinical studies that normally cannot be carried out in the target population; these studies are often enough on their own to establish correlation. By also establishing mechanism, one can then establish causality in the study population. In the target population, clinical studies in the target population, even when augmented with other evidence of the mechanisms of the target population, are insufficient to establish correlation and mechanism—otherwise there would be no need for extrapolation. Extrapolation is possible when evidence of mechanisms in the target population is strong enough not only to establish the *existence* of a suitable mechanism M' in the target population, but also to establish that this mechanism is *similar* in key respects to the mechanism M inferred in the study population. The expression $M' \equiv M$ in Fig. 12 denotes this similarity claim. By means of this similarity of mechanisms, one can use the claim that A is a cause of B established in the study population to further support the correlation claim in the target population. In sum, where clinical studies and other mechanistic investigations in the target population are not jointly sufficient to establish correlation in the target, if the corresponding causal claim is established in the study population and it is also established that the mechanisms in the target population are sufficiently similar to those which underpin causation in the study population then this combination of evidence may be enough to establish correlation in the target population. If so, since mechanism in the target is also established, causality can be inferred.

As an extreme case, there may be no clinical studies in the target population; this in itself does not preclude extrapolation under the RWT-motivated account. For example, when IARC evaluated benzo[a]pyrene, they found no human studies measuring exposure to benzo[a]pyrene together with relevant cancer outcomes. However, there were excellent animal studies and enough evidence of mechanisms

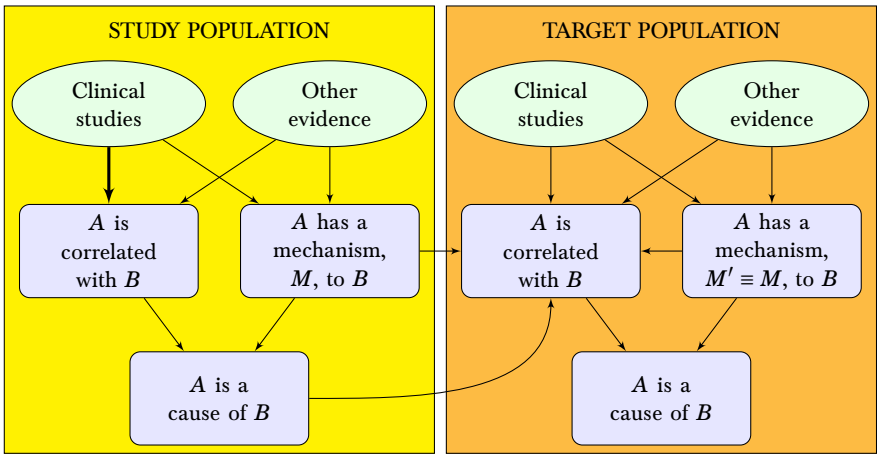


Figure 12: The logic of extrapolation as motivated by RWT.

in animals to establish carcinogenicity in the relevant animal models and to determine the details of the mechanism of action there. Furthermore, there was excellent evidence that the human mechanisms were similar to the mechanisms found in animals. This was considered enough to establish carcinogenicity in humans (IARC, 2012, pp. 111–144).[3] Note that this example is not validated by the EBM-motivated account of extrapolation provided above, because there were no relevant clinical studies in humans. Thus the example favours the RWT-motivated account of extrapolation.

To take another case where there were no clinical studies in the target population, consider the IARC evaluation of *d*-Limonene as a cause of cancer. In this case, there were no studies available in humans. Carcinogenicity of *d*-Limonene was established in male rats, so this seemed to be a candidate for extrapolation. However, there were crucial dissimilarities between the mechanism of action in rats and the corresponding human mechanisms: in particular, a protein responsible for nephrotoxicity in male rats is specific to male rats. Thus no extrapolation was possible and carcinogenicity was not established (IARC, 1999, pp. 3017–327).[4] This example, which is also in accord with the RWT-motivated account, shows how crucial it is to establish similarity of mechanisms.

⋆

We shall now see that this RWT-motivated account of extrapolation survives the four problems of extrapolation identified above.

First let us consider the extrapolator's circle. That there is no circle should be apparent from the fact that Fig. 12 is acyclic: one does not need to have already established causality in the target in order to meet any of the requirements of establishing causality. Of course, once these requirements are all met, causality in the target is thereby established, but there is no inferential circle here. See Steel

[3] I am very grateful to Michael Wilde for alerting me to this example.

[4] I am very grateful to Kurt Straif for alerting me to this example.

(2008, §5.4.2) for further discussion of how mechanism-based approaches can avoid the extrapolator's circle.

Turning next to the extrapolator's block, one might worry that we are lacking an account of how extrapolation is possible when mechanisms in the study and target populations are not identical. Similarity of mechanisms is a matter of degree, and the more similar the mechanisms, the more causation in the study population confirms correlation in the target population. Steel (2008, §5.3.2) discusses this question and presents *comparative process tracing* as a method for establishing similarity:

> First, learn the mechanism in the model organism, by means of process tracing or other experimental means. For example, a description of a carcinogenic mechanism would indicate such things as the product of the phase I metabolism and the enzymes involved; whether the metabolite is a mutagen, an indication of how it alters DNA; and so on. Second, compare stages of the mechanism in the model organism with that of the target organism in which the two are most likely to differ significantly. For example, one would want to know whether the chemical is metabolized by the same enzymes in the two species, and whether the same metabolite results, and so forth. In general, the greater the similarity of configuration and behavior of entities involved in the mechanism at these key stages, the stronger the basis for the extrapolation. (Steel, 2008, p. 89.)

Two further points are important here. First, it is important to show that the whole *structure* of relevant mechanisms is sufficiently similar, not just that the mechanism M by which causality operates in the study population has an analogue in the target population. Thus, one needs to establish that any new counteracting mechanisms in the target population are not so significant that they can cancel out the action of the analogue of M. Second, it is important to note that comparative process tracing is but one of several methods for establishing similarity of mechanisms. One can also establish similarity of mechanisms without determining the details of the mechanisms M and M' by employing phylogenetic reasoning, robustness analysis or even enumerative induction (Parkkinen and Williamson, 2017, §4).

Let us consider the extrapolator's fallacy next. Unlike the EBM-motivated approach, the RWT-motivated analysis of extrapolation requires evidence that ensures that the study and target populations are similar in causally relevant respects. Mechanistic evidence is playing a key role here, in ensuring that $M' \equiv M$. By being more demanding in terms of the evidence required in the target population, extrapolation promises to be more reliable under the RWT account than under the EBM account.

Finally, we can ask whether the extrapolator's standards are slipping. That this is not so is apparent from Fig. 12: the inferential requirements—establishing correlation and mechanism—are the same in both the study and target populations. If anything, one might one worry that the standards of evidence are higher in the target population than in the study, since Fig. 12 includes the extra requirement of establishing similarity of mechanism there. However, this is just an artefact of the diagram. Similarity of mechanisms concerns the relation between the study and target populations, not just the target population. Therefore, there is a genuine symmetry between what is required of the study and target populations.

That the RWT account of extrapolation overcomes the latter two problems while the EBM approach does not, speaks in favour of the RWT approach.

⋆

Having developed the RWT-motivated theory of extrapolation, we shall now consider some criticisms of mechanistic accounts of extrapolation in the light of this theory.

Guala (2010, §6) suggests that there are cases of extrapolation that do not proceed via comparative process tracing. Guala develops an example involving outer continental shelf auctions, which are used to sell oil leases in the Gulf of Mexico, to show that it is not always necessary to determine the details of the relevant mechanisms, as would be required by comparative process tracing. As noted above, the RWT-motivated account sees comparative process tracing as but one of several strategies for establishing similarity of mechanisms, and Guala's case is perfectly in accord with this. What is important to the RWT account is the inferential step $M' \equiv M$: strategies for extrapolation seek to demonstrate *similarity of mechanisms*. As Guala notes,

> This clearly falls short of a proper articulation of the mechanism … And yet, it is perfectly adequate for extrapolation purposes. Large parts of the mechanism can be "black boxed" as long as there are good reasons to believe that they are analogously instantiated in the laboratory and target system.' (Guala, 2010, p. 1080.)

One of the advantages of the RWT-motivated approach, then, is that by situating extrapolation in the inference scheme depicted by Fig. 12 it covers much a broader range of scenarios than comparative process tracing.

Howick et al. (2013a,b) are broadly sceptical of mechanism-based extrapolation. They identify several problems for basing extrapolations on mechanistic evidence. First, our understanding of mechanisms is often incomplete. This is of course true, but insufficient knowledge of the details of M and M' for comparative process tracing does not always preclude establishing that $M' \equiv M$: one can often employ the other strategies discussed above. Second, knowledge of mechanisms is not always applicable outside the tightly controlled laboratory conditions in which is gained. This is also true, but it is symptomatic of science in general: whatever approach one takes, one must make sure that one's conclusions are robust enough to extend to the application of interest. In particular, an EBM-motivated approach has to ensure that conclusions based on trials with strict exclusion criteria are transportable to the population to be treated. The third problem that they identify is that mechanisms can behave 'paradoxically', e.g., a drug can have opposite effects in different contexts. However, it is only by understanding the underlying mechanisms that one can explain these paradoxical effects and improve treatment. On the other hand, clinical studies are crucial for identifying such effects, and one advantage of the RWT-motivated account is that it is not exclusively mechanism-based: it treats evidence of correlation and evidence of mechanisms on a par. The fourth problem that they pick out is the extrapolator's circle. Their worry is that the evidence of the target population required to establish that $M' \equiv M$ makes the evidence on the study population redundant. As Fig. 12 makes clear, this need not be the case: one can establish that $M' \equiv M$ in the absence of evidence from clinical studies in the target population that would be sufficient to establish causality. Howick et

al. might respond by noting that under the EBM-motivated account of extrapolation presented above, only weak evidence of the target population is required to establish causality in the target population. However, as discussed above, this is a problem for the EBM-motivated account: it makes extrapolations too easy to be entirely credible—it is subject to the extrapolator's fallacy. That the RWT-motivated theory of extrapolation is more demanding in terms of the evidence required for extrapolation is an advantage over the EBM-motivated account.

§4 Conclusion

We have seen that the epistemological thesis RWT motivates a view of medical methodology that stands in conflict with contemporary EBM. Although there is a tension between RWT and EBM, I have argued that RWT can better explain three key features of the use of clinical studies to establish causality, and that it yields a better account of extrapolation. Thus, I conclude that EBM+ is a promising way forward in the controversy as to how best to improve evidence based medicine.

The EBM approach to causal inference has in recent years extended well beyond medicine, to public policy making and various areas of the social sciences, for example. While this paper has focussed on medicine, RWT can be interpreted as having a broader range of application, and similar conclusions to those drawn in this paper may apply beyond medicine. The broader scope of these conclusions is left as a question for further research.

Acknowledgements

I am very grateful to the Leverhulme Trust and to the UK Arts and Humanities Research Council for supporting this research.

Bibliography

Aronson, J. K. and Hauben, M. (2006). Anecdotes that provide definitive evidence. *British Medical Journal*, 333(7581):1267–1269.

Bartlett, R., Andrews, A., Toomasian, J., Haiduc, N., and Gazzaniga, A. (1982). Extracorporeal membrane oxygenation for newborn respiratory failure: forty-five cases. *Surgery*, 92(2):425Ů433.

Bartlett, R. H., Roloff, D. W., Cornell, R. G., Andrews, A. F., Dillon, P. W., and Zwischenberger, J. B. (1985). Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study. *Pediatrics*, 76(4):479–487.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100:407–425.

Campaner, R. (2011). Understanding mechanisms in the health sciences. *Theoretical Medicine and Bioethics*, 32:5–17.

Campaner, R. and Galavotti, M. C. (2012). Evidence and the assessment of causal relations in the health sciences. *International Studies in the Philosophy of Science*, 26(1):27–45.

Clarke, B. (2011). *Causality in medicine with particular reference to the viral causation of cancers*. PhD thesis, Department of Science and Technology Studies, University College London, London.

Clarke, B., Gillies, D., Illari, P., Russo, F., and Williamson, J. (2013). The evidence that evidence-based medicine omits. *Preventative Medicine*, 57(6):745–747.

Clarke, B., Gillies, D., Illari, P., Russo, F., and Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi*, 33(2):339–360.

Claveau, F. (2012). The Russo-Williamson theses in the social sciences: Causal inference drawing on two types of evidence. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(4):806–813.

Cucherat, M., Haugh, M. C., Gooch, M., and Boissel, J.-P. (2000). Evidence of clinical efficacy of homeopathy: A meta-analysis of clinical trials. *European Journal of Clinical Pharmacology*, 56:27–33.

Darby, G. and Williamson, J. (2011). Imaging technology and the philosophy of causality. *Philosophy & Technology*, 24(2):115–136.

Dragulinescu, S. (2012). On 'stabilising' medical mechanisms, truth-makers and epistemic causality: a critique to Williamson and Russo's approach. *Synthese*, 187(2):785–800.

Faculty of Homeopathy (2016). The research evidence base for homeopathy. http://facultyofhomeopathy.org/wp-content/uploads/2016/03/2-page-evidence-summary-for-homeopathy.pdf.

Fiorentino, A. R. and Dammann, O. (2015). Evidence, illness, and causation: An epidemiological perspective on the Russo-Williamson thesis. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 54:1–9.

Gillies, D. A. (2011). The Russo-Williamson thesis and the question of whether smoking causes heart disease. In Illari, P. M., Russo, F., and Williamson, J., editors, *Causality in the Sciences*, pages 110–125. Oxford University Press, Oxford.

Glasziou, P., Chalmers, I., Rawlins, M., and McCulloch, P. (2007). When are randomised trials unnecessary? Picking signal from noise. *British Medical Journal*,

334(7589):349–351.

Gruta, N. L. L. and Turner, S. J. (2014). T cell mediated immunity to influenza: mechanisms of viral control. *Trends in Immunology*, 35(8):396–402.

Guala, F. (2010). Extrapolation, analogy, and comparative process tracing. *Philosophy of Science*, 77:1070–1082.

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., deBeer, H., Jaeschke, R., Rind, D., Meerpohl, J., Dahm, P., and Schünemann, H. J. (2011). {GRADE} guidelines: 1. introductionŮgrade evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, 64(4):383–394.

Hauben, M. and Aronson, J. K. (2007). Gold standards in pharmacovigilance: The use of definitive anecdotal reports of adverse drug reactions as pure gold and high-grade ore. *Drug Safety*, 30(8):645–655.

Howick, J. (2011a). Exposing the vanities—and a qualified defence—of mechanistic evidence in clinical decision-making. *Philosophy of Science*, 78(5):926–940. Proceedings of the Biennial PSA 2010.

Howick, J. (2011b). *The philosophy of evidence-based medicine*. Wiley-Blackwell, Chichester.

Howick, J., Glasziou, P., and Aronson, J. (2013a). Can understanding mechanisms solve the problem of extrapolating from study to target populations (the problem of 'external validity')? *Journal of the Royal Society of Medicine*, 106(3):81–86.

Howick, J., Glasziou, P., and Aronson, J. K. (2013b). Problems with using mechanisms to solve the problem of extrapolation. *Theoretical Medicine and Bioethics*, 34(4):275–291.

IARC (1999). *Some Chemicals that Cause Tumours of the Kidney or Urinary Bladder in Rodents and Some Other Substances*, volume 73 of *IARC Monographs Series*. International Agency for Research on Cancer, Lyon. http://monographs.iarc.fr/ENG/Monographs/vol73/mono73.pdf.

IARC (2002). *Some Traditional Herbal Medicines, Some Mycotoxins, Naphthalene and Styrene*, volume 82 of *IARC Monographs Series*. International Agency for Research on Cancer, Lyon. http://monographs.iarc.fr/ENG/Monographs/vol100A/mono100A.pdf.

IARC (2012). *Chemical Agents and Related Occupations*, volume 100F of *IARC Mongraphs Series*. International Agency for Research on Cancer, Lyon. http://monographs.iarc.fr/ENG/Monographs/vol100F/mono100F.pdf.

IARC (2013). *Non-ionizing radiation, part 2: radiofrequency electromagnetic fields*, volume 102 of *IARC Mongraphs Series*. International Agency for Research on Cancer, Lyon. http://monographs.iarc.fr/ENG/Monographs/vol102/mono102.pdf.

Illari, P. M. (2011). Disambiguating the Russo-Williamson thesis. *International Studies in the Philosophy of Science*, 25(2):139–157.

Illari, P. M. and Williamson, J. (2012). What is a mechanism? thinking about mechanisms *across* the sciences. *European Journal for Philosophy of Science*, 2:119–135.

Ioannidis, J. P. A. (2012). Extrapolating from animals to humans. *Science Translational Medicine*, 4(151):151ps15.

Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge University Press, Cambridge.

Leibovici, L. (2001). Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial. *British Med-*

ical Journal, 323:1450–1451.

Machamer, P., Darden, L., and Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science*, 67:1–25.

Mugford, M., Elbourne, D., and Field, D. (2010). Cochrane review: Extracorporeal membrane oxygenation for severe respiratory failure in newborn infants. *Evidence-Based Child Health: A Cochrane Review Journal*, 5(1):241–298.

NICE (2006). *The guidelines manual*. National Institute for Health and Clinical Excellence, London. Available from: www.nice.org.uk.

OCEBM Levels of Evidence Working Group (2011). The Oxford 2011 levels of evidence. Oxford Centre for Evidence-Based Medicine, http://www.cebm.net/index.aspx?o=5653.

Parkkinen, V.-P. and Williamson, J. (2017). Extrapolating from model organisms in pharmacology. In Osimani, B., editor, *Uncertainty in pharmacology: epistemology, methods, and decisions*. Springer, Dordrecht.

Reichenbach, H. (1956). *The direction of time*. University of California Press, Berkeley and Los Angeles, 1971 edition.

Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.

Russo, F. and Williamson, J. (2011a). Epistemic causality and evidence-based medicine. *History and Philosophy of the Life Sciences*, 33(4):563–582.

Russo, F. and Williamson, J. (2011b). Generic versus single-case causality: the case of autopsy. *European Journal for Philosophy of Science*, 1(1):47–69.

Sackett, D. L., Rosenberg, W. M. C., Gray, J. A. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal*, 312(7023):71–72.

Salmon, W. C. (1998). *Causality and explanation*. Oxford University Press, Oxford.

Smith, G. C. S. and Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *British Medical Journal*, 327:1459–1460.

Steel, D. (2008). *Across the Boundaries, Extrapolation in Biology and Social Science*. Oxford University Press, Oxford.

SUNY (2004). SUNY Downstate Medical Center evidence based medicine tutorial. http://library.downstate.edu/EBM2/contents.htm. The Medical Research Library of Brooklyn.

Weber, E. (2007). Social mechanisms, causal inference, and the policy relevance of social science. *Philosophy of the Social Sciences*, 30(3):348–359.

Weber, E. (2009). How probabilistic causation can account for the use of mechanistic evidence. *International Studies in the Philosophy of Science*, 23(3):277–295.

Williamson, J. (2009). Probabilistic theories. In Beebee, H., Hitchcock, C., and Menzies, P., editors, *The Oxford Handbook of Causation*, pages 185–212. Oxford University Press, Oxford.

Williamson, J. (2015). Deliberation, judgement and the nature of evidence. *Economics and Philosophy*, 31(1):27–65.

Worrall, J. (2007). Why there's no cause to randomize. *British Journal for the Philosophy of Science*, 58:451–488.

Imprecise probability and biological fitness

Word count: 5000

March 1, 2016

Abstract

I argue that biological fitness sometimes depends on imprecise probabilities. I give a new argument that some outcomes are without objective probability, and argue that organisms encountering environments might sometimes be outcomes of this kind. I argue that since fitness depends on relationships between traits and environments, this means that fitness can depend on imprecise probabilities, and can be defined by an interval between maximum and minimum precise fitnesses. One trait is fitter than another when its minimum fitness is greater than the other's maximum fitness or when, in some conditions, the first trait has greater fitness in every environment.

1 Introduction

Despite numerous connections between evolutionary biology and decision theory (e.g. Alexander 2007; Okasha 2007; Wagner 2010; Okasha and Binmore 2012), work on decision theory with imprecise probabilities (e.g. Levi 1980; Walley 1991; Joyce 2010; Moss 2015) has had little if any impact on discussions of fitness. Drawing upon ideas from decision making with imprecise probability, I argue that concepts of biological fitness, which have traditionally been defined in terms of probability, should be extended to allow definition in terms of imprecise probabilities. I explore the resulting notions of what I call

"imprecise fitness", focusing especially on the possibility of cases in which environments vary in ways that can't be captured by objective probability. Using using a new argument that some outcomes are without objective probability, I argue that such cases of *erratic* environmental variation plausibly exist. One trait is then fitter than another if the least fitness of the first is larger than the greatest fitness of the other's, or if the weaker relation of being fitter in every environment holds between them in certain situations.

In section 2 I introduce the idea of imprecise probability. Section 3 argues that some outcomes occur *erratically*, i.e. according to no objective probability. This argument, which is not specific to evolutionary contexts, suggests a more general role for objective imprecise probability than has previously been proposed. Section 4 introduces ideas about fitness and environmental variation that I use as a starting point for arguments in section 5 that some fitnesses are imprecise. This section also discusses senses in which one trait can be fitter than another when their fitnesses are imprecise. Section 6 summarizes my conclusions and remarks on their relationship to work on imprecise decision theory.

2 Imprecise probability

A probability measure assigns precise, real-valued probabilities to a set of outcomes. By contrast, an imprecise probability function associates a set of probability measures with a set of outcomes. For each each outcome, there is then a set of real-valued probabilities. Imprecise probabilities are usually discussed as an extension of Bayesian credence (e.g. Walley 1991; Joyce 2010; Troffaes and de Cooman 2014; Moss 2015; Schoenfield 2012; Bradley 2015b; Bradley and Steele 2016). For example, we might think of a particular person's degree of belief that P as interval valued—representable by a pair of numbers, rather rather than a single number—where the two numbers are the minimum and maximum values assigned by those probability measures that represent the person's system of beliefs. Some authors have also argued that conceptions of objective imprecise probabil-

ity may be useful. For example, Dardashti et al. (2014) argued that Best System chances (Lewis, 1994) may be imprecise, because it might not be determinate which system of laws best trades off simplicity, strength, and fit. Hajék and Smithson (2012) argued that the existence of real-world cases like one described in (Papamarcou and Fine, 1986) might mean that Best System chances should be imprecise.

Terrence Fine and his collaborators have argued that imprecise probabilities may sometimes be features of the world (e.g. Walley and Fine 1982; Papamarcou and Fine 1986; Fierens et al. 2009; Fierens 2009). This view does not depend on Best System considerations, but starts from the idea that some objective probabilities are long-run frequencies, generalizing it to sequences in which frequencies don't tend toward a limit. For example, Fierens et al. (2009) explored the idea that in some cases there is a process that determines, for each trial in a long but finite sequence, which of several objective probability distributions applies to the outcome of that trial. This process then defines imprecise probabilities if the probability-choosing process is such that the sequence of probability distributions has an intermediate level of Kolmogorov complexity.[1] Note that the per-trial probabilities must be objective, and Fierens et al. (2009) suggest that they are propensities, but the framework is consistent with other ways of defining objective probabilities.

[1]If the Kolmogorov complexity is high, there's a probability distribution P that governs which individual probability distribution P_k is chosen for each trial (e.g. Li and Vitányi 2008, chapter 2). An outcome then has a precise probability that is a P-weighted average of trial probabilities P_k for the outcome. If the Kolmogorov complexity is low, the process that determines per-trial probabilities P_k is very simple—for example, alternating between two distributions every 10 trials—so it makes more sense to say that we have an alternation of (precise) probability distributions. Imprecise probabilities arise when the alternation of probability distributions on trials is neither as regular as the latter kind of case nor as systematically random as the former.

3 Objectively erratic outcomes

Any objective probability[2] is a probability of an outcome—of a set of *occurrences* (token events), of the realization of a type by an occurrence, or of a proposition—in a space of alternatives. Even the probability that this coin on this toss will trace exactly this path and land heads in this location is a probability of a type, for it's multiply realizable. Now consider mutually exclusive occurrences e_1 and e_2 (e.g. a particular dollar coin being at position p_1 on a table at time t, or instead being at position p_2 at t). If these two occurrences have objective probabilities, they do so "under some description". It is as realizations of properties that e_1 and e_2 have objective probabilities. Must there be objective probabilities of the coin *being at position* p_1 *on the table* or *being at position* p_2 *on the table*? I don't see why there must be. Suppose there were probabilities of the coin being given to one of two children, Araceli and Billy, in either the pile of gifts for Araceli (which may happen to be at p_1) or the pile for Billy (which may be at p_2). Must there also be a probability that the piles are in particular places? That all coins in Araceli's pile are free of nicks on their edges? That at t the coin is within a mile of someone who is over seven feet tall? Why should all possible properties of the coin have probabilities? Even if it were true that every occurrence has a probability as a realizer of *some* property, this doesn't imply that every algebra over a set of possible, mutually exclusive occurrences defines a space of outcomes with objective probabilities. Even in a deterministic world, where any token event that occurs had to occur, whether that occurrence has a probability is relative to a set of possible properties.[3]

I conclude that some outcomes have no probabilities. We can refer these outcomes as

[2] I mean probabilities that in some sense cause and can be used to manipulate frequencies, as in (Abrams, 2015).

[3] Outcomes realized only by effects of deterministic processes may have objective probabilities other than 0 and 1—for example "Spielraum probabilities" (Beisbart, online January 2016; Rosenthal, 2012; Strevens, 2011; Abrams, 2012).

occurring *erratically* (Hájek and Smithson, 2012), or with *erraticity*. This doesn't mean that token events realizing erratically determined outcomes happen for no reason. They are caused, either deterministically, or indeterministically in the way that quantum mechanical events are thought to occur. Erratically occurring outcomes have relative frequencies in practice, but there is no predictability to their frequencies except in the sense that particular occurrences that realize the outcomes may be predictable. Another set of occurrences that realize outcomes in the same space may have radically different frequencies, even if they are produced in a similar way.

4 Fitness and environmental variation

A primary theoretical role of fitness is in defining natural selection. Natural selection involves, a the very least, (possibly null) changes in relative frequencies of phenotypic traits, genotypes, or alleles present in a population over time; fitness differences play a crucial role in this process, explaining and potentially predicting such changes. Any attempt to make sense of the theoretical role of fitness must provide some account of the fitnesses of traits—in particular, heritable traits, since it's only those whose frequencies natural selection can affect over many generations (cf. Lewontin 1970; Godfrey-Smith 2009). Though many authors define a notion of fitness that has a value for each actual token organism (e.g. Mills and Beatty 1979; Brandon 1990; Pence and Ramsey 2013), it will be fruitful to focus on trait fitnesses.

Most conceptions of fitness assume, at least tacitly, that there are probabilities of biological outcomes such as having certain numbers of offspring. The idea that trait fitness is an average number of offspring for actual or possible organisms with the trait is common in evolutionary biology (e.g. Stearns 1992), but it's become clear that defining fitness as probability-weighted expected number of offspring is not always appropriate (Gillespie, 1977; Levins, 1968; Beatty and Finsen, 1989; Brandon, 1990; Sober, 2001; Abrams, 2009a).

Nevertheless, I'll restrict my discussion to fitness as expectation. In some cases, we can simply generalize a definition of fitness as expected number of offspring to expected numbers of descendants at some particular later generation.

Fitness is always relative to an environment: Were a population of organisms placed in a different environment (a cold rather than warm one, for example), the relative fitnesses of traits might differ. It's common to extend this idea to smaller "environments" (habitats, patches, subenvironments) within an overall environment. These variant environments may be arranged spatially within the overall environment, or they may occur at different times, or both. Consider an example like Abrams' (2009b) or Walsh's (2013): In a population of animals, one trait, *deep*, leads its bearers to dig deep burrows, while the other, *shallow*, leads its bearers to dig shallow burrows. The *deep* trait makes drowning more likely during periods of torrential rain, but it is better than *shallow* during hot, dry periods because cool burrows are advantageous. Suppose that the fitness of *deep* is 1 during rainy ("wet") periods and 2 during dry periods, while the fitness of *shallow* is 2 during wet periods and 1 when it is dry. Which trait is fitter? Which will natural selection most likely favor? If the *wet* environment is more probable, then *shallow* is fitter, and vice versa if the *dry* environment is more probable. Overall fitness w can be defined relative to the overall environment, which is defined, in part, by the probabilities of the *wet* and *dry* environments. For example, if $\mathsf{w}(d)_{wet}$ is the fitness of *deep* in the *wet* environments, and $\mathsf{w}(d)_{dry}$ is the fitness of *deep* in *dry* environments, then the overall fitness of *deep* might[4] be $\mathsf{w}(d)_{wet} \times \mathsf{P}(wet) + \mathsf{w}(d)_{dry} \times \mathsf{P}(dry)$.

(Given a set of mutually exclusive, competing heritable traits, we can view other dimensions of trait variation as defining "environments". For example, an allele—a gene—at a genetic locus on a chromosome is a kind of minimal heritable trait, and it may be copied more or less often into "environments" consisting alleles at other loci within the genome. More generally, we can view heritable phenotypic traits as environmental contexts for

[4]See (Levins, 1968; Abrams, 2014).

other heritable traits when they may be present in the same organisms.)

5 Imprecise fitness

Imprecise fitness It's often assumed that that there are determinate probabilities of environmental fluctuations, or of organisms finding themselves in different environments (e.g. Levins 1968; Brandon 1990; Gillespie 1998; Ramsey 2006; Abrams 2009b), but this assumption doesn't seem required. There may be some cases in which some environmental changes that affect evolutionary success obey no probabilities: Though in some particular period of time, organisms would encounter some environments more often than others, there would be no reason for this pattern, or any pattern, to be likely to continue over many generations.[5]

I believe that empirical research in evolutionary biology shows that statistical methods and modeling in terms of probabilities concerning evolutionary outcomes have been enormously useful. This extends to models of environmental variation (e.g. Giacomini and Shuter 2013), suggesting that there are in fact objective probabilities that organisms with given traits encounter particular environmental states (Abrams, 2015). However, successful use of probability in such methods needn't imply that all evolutionary processes are governed only by objective precise probabilities. Even successful probabilistic models may owe their success to objective imprecise probabilities that are easily approximated with precise probabilities in models.

When there are no probabilities that organisms will encounter particular environments, can we ever say that one trait is fitter than another? Yes. For example, if the expected numbers of offspring for *deep* in both *wet* and *dry* is greater than that of *shallow* in

[5]Nevertheless these environments e_i might consist of smaller or more narrowly defined, mutually exclusive environments e_{ij}, where part of what defines e_i are conditional probabilities $\mathrm{P}(e_{ij}|e_i)$ of its component environments e_{ij} occurring. The precise fitness of a trait in e_i can then be defined like overall fitness, but using probabilities conditional on e_i.

either environment, then whatever the frequency with which *wet* and *dry* alternate, *deep* will have a higher average number of offspring. Suppose, for example, that the expected numbers of offspring for *deep* are 1.5 and 1.7 in *wet* and *dry* environments, respectively, and the corresponding numbers for *shallow* are 1.2 and 0.9. Even if the environment is, as it happens, erratically, always in the *wet* state—the one that gives *shallow* its higher expected number of offspring—*deep* should usually have a greater number of offspring on average than *shallow*.

Such cases can be represented by treating fitness of a trait A as a pair of numbers representing the minimum $\underline{\mathrm{w}}(A)$ and maximum $\overline{\mathrm{w}}(A)$ precise fitnesses that the trait has in the environments that organisms erratically encounter. Here $\underline{\mathrm{w}}$ and $\overline{\mathrm{w}}$ are *lower prevision* and *upper prevision* operators, respectively; these are generalizations of expectation for imprecise probability (Walley, 1991; Troffaes and de Cooman, 2014; Augustin et al., 2014).[6] I'll refer to such a pair a minimum and maximum fitnesses—the values of particular lower prevision and upper prevision operators applied to a trait—as a "fitness interval", represented by the notation $[l, h]$, where l and h are lower and upper previsions, respectively, for the trait. In the *wet/dry* example, it's reasonable to call the pair of fitness values for a trait a fitness "interval", since what happens in the world will involve some combination of the two environmental states. If organisms in the population always found themselves in just one of the two environments, then the expected number of offspring for the trait would be either l or h. Otherwise, the result would be a weighted average of l and h, but this average would depend on what environments were actually encountered (erratically) by organisms in the population. More generally, for more than two erratically determined environmental states generating different fitness values, the fitness of the trait is

[6]So defined, the $\underline{\mathrm{w}}$ and $\overline{\mathrm{w}}$ functions are *coherent* lower and upper previsions, by the lower envelope theorem (Troffaes and de Cooman, 2014, Theorem 4.38, p. 71). To say that lower and upper previsions are coherent means that that they satisfy certain constraints involving linear combinations of gambles (paid off in numbers of offspring, here) and lower or upper previsions of those gambles (Walley, 1991; Troffaes and de Cooman, 2014; Augustin et al., 2014).

still captured by the interval between its minimum and maximum precise fitnesses across all of the environments that are possible in the overall environment.

Interval dominance The example above in which the fitness of *deep* is $[1.5, 1.7]$ and that of *shallow* is $[0.9, 1.2]$ illustrates a general claim that can expressed in terms of the *interval dominance* relation, $\sqsupset$ (Troffaes, 2007; Huntley et al., 2014; Bradley, 2015a):[7]

$$A_1 \sqsupset A_2 \quad \text{iff} \quad \underline{\mathrm{w}}(A_1) > \overline{\mathrm{w}}(A_2).$$

That is, $A_1 \sqsupset A_2$ iff A_1's minimum fitness is greater than A_2's maximum fitness (across all environments). We can also say that A_1's *lower fitness* $\underline{\mathrm{w}}(A_1)$ is greater than A_2's *upper fitness* $\overline{\mathrm{w}}(A_2)$. This gives us a partial ordering of traits:

A_1 is fitter than A_2 if $A_1 \sqsupset A_2$.

A_1 is fittest if $A_1 \sqsupset A_j$ for all competing traits A_j.

This is a partial order because it may be that neither A_1 or A_2 is fitter than the other in this sense, because the two traits' fitness intervals overlap, i.e. when $\underline{\mathrm{w}}(A_1) \leq \overline{\mathrm{w}}(A_2)$ and $\underline{\mathrm{w}}(A_2) \leq \overline{\mathrm{w}}(A_1)$.

Dominance across population-wide environments There is at least one kind of case in which a trait can be considered fitter than another though they have overlapping fitness intervals. Suppose that trait A_1 dominates trait A_2 in the sense that in every environment e, the precise fitness $\mathrm{w}_e(A_1)$ of A_1 relative to that environment is greater the corresponding precise fitness $\mathrm{w}_e(A_2)$ for A_2 (cf. Bradley 2015a). It's then probable that the instances of trait A_1 will have a higher average number of offspring than instances of A_2 *if* the environmental variation is such that all members of the population experience the same environment at any given time, as in the case of population-wide temporal variation in environments. More precisely, there must be generations that don't overlap, as when

[7] The interval dominance relation is used in (Troffaes, 2007; Huntley et al., 2014) to define a somewhat different rule for choosing gambles, also called "interval dominance".

organisms lay eggs and then die (although this needn't happen in every generation). The idea is that environments can change erratically only between two non-overlapping generations. In such cases say that A_1 *dominates* A_2 *across population-wide environments*.

For other kinds of environmental variation, that A_1 dominates A_2 in every environment need not imply that A_1 will probably increase in frequency. For example, suppose that the environment of A_1 and A_2 is composed of two spatially varying environments e_1 and e_2, and that whether a given organism ends up in one environment or the other is merely erratic. (The organisms might be plants whose seeds are distributed by the wind to soil patches of two kinds, if those aspects of the wind that affect seed distribution are erratic.) Assume that A_1 has precise fitness $\mathrm{w}_{e_1}(A_1) = 1.5$ in environment e_1, and fitness $\mathrm{w}_{e_2}(A_1) = 2.5$ in environment e_2, while corresponding fitnesses for A_2 are $\mathrm{w}_{e_1}(A_2) = 1$ and $\mathrm{w}_{e_1}(A_2) = 2$. Thus A_1 dominates A_2 in every environment. However, since it's erratic which tokens of A_1 or A_2 are to be found in either environment, it may turn out that most of the A_2's end up in e_2 where they have a fitness of 2, while most of the A_1's end up in e_1, where their fitness is 1.5. Given that actual distribution of A_1's and A_2's it would be probable that A_2 would increase in frequency even though A_1 dominates A_2 (but that distribution is neither probable nor improbable).

Summarizing the preceding points:

> If environments e vary erratically in such a way that any time, the entire population experiences the same environment, then:
> A_1 is fitter than A_2 if $(\forall e)\mathrm{w}_e(A_1) > \mathrm{w}_e(A_2)$.
> A_1 is fittest if $(\forall e)\mathrm{w}_e(A_1) > \mathrm{w}_e(A_j)$ for all competing traits A_j.[8]

(If we change the strict strict inequality to $\geq$, we can replace "fitter" and "fittest" with "is at least as fit as" and "is among the fittest", respectively. An analogous generalization of interval dominance is not so straightforward.)

[8] Cf. (Brandon, 1990, chapter 2) and (Abrams, 2014) on "selective environments", Joyce's (2010) choice function 4I, and Rinard's (2015) Moderate choice function.

Fitter than Note that if A_1 interval dominates A_2, then A_1 also dominates A_2 in every environment, since $A_1 \sqsupset A_2$ means that the lowest precise fitness of A_1 in any environment is greater than the highest fitness of A_2 in any environment. Thus we can summarize the two kinds of fitness relation above as follows:

> A_1 is fitter than A_2 if and only if either $A_1 \sqsupset A_2$, or all organisms experience the same environment e at the same time and $(\forall e)\mathrm{w}_e(A_1) > \mathrm{w}_e(A_2)$.

This generalizes traditional meaning of "fitter than" for fitnesses that are expected numbers of offspring. Note that it may be that there is a set $\mathscr{A}$ of traits A_i such that each A_i is fitter than all traits not in in $\mathscr{A}$, but that no trait in $\mathscr{A}$ is fitter than any other in $\mathscr{A}$.

If A_1 is not fitter than A_2 in either of the preceding senses, it would be misleading to say that they are equal in imprecise fitness, since that would suggests that the traits' evolutionary successes would usually be similar, at least in the short run, when there are many organisms with those traits. It's better to say that when neither A_1 nor A_2 is (imprecisely) fitter than the other, the two traits are *incomparable*, or that it's *indeterminate* (cf. Rinard 2015) which is fitter.

E-admissibility It's not clear that anything more can be said about fitness inequalities for imprecise fitness. For example, consider this adaptation to fitness of Levi's (1980) *E-admissibility* rule for choosing gambles:[9]

> The fittest traits are those A_i for which there's some environment in which A_i's precise fitness is a least as great as that of any trait A_j.

If organisms with such a trait A_i and those without such a trait both encounter an environment in which A_i's precise fitness is greater, it's likely that those with A_i will have more offspring, on average, than the others. The problem is that if the environments encoun-

[9]The decision theoretic rule requires that acceptable gambles must have the highest expected value according to at least one of the decision maker's credence distributions (Levi, 1980; Huntley et al., 2014). Moss (2015) calls this rule "permissive"; White (2009) and Rinard (2015) call it "Liberal".

tered by organisms are determined erratically, there's no reason that organisms of both kinds would encounter such an environment often—or ever. Thus the E-admissibility of traits provides no information about evolutionary success.

6 Conclusion

I argued that some outcomes are determined erratically, i.e. according to no objective probabilities, and that it's plausible that some environmental variation is erratic. (More generally, there may be reasons other than erratic environmental variation that probability distributions relevant to evolutionary success are erratically determined.) Where environmental variation is erratic, fitnesses can be understood as intervals between a *lower fitness*—the minimum precise fitness over all erratically-determined environments—and an *upper fitness*—the corresponding maximum. We can say then that a trait A_1 is fitter than another trait A_2 when A_1 is interval dominant over A_2—when A_1's lower fitness is greater than A_2's upper fitness—or when organisms in the same generation are always subject to the same environment, and A_1's precise fitness is greater than A_2's in every environment. In other cases, it appears that fitness intervals are incomparable; they neither predict nor explain A_1 or A_2's ultimate success.

It may be that some lack of fit between models and evolutionary processes is due to objective imprecise probabilities. Nevertheless, if fitness intervals are often narrow, it would be useful to model natural selection in terms of probabilities and precise fitnesses.

Analogies between my conclusions and well known positions on decision making with imprecise probabilities are not as close as one might have expected. However, there are disanalogies between evolution and decision making. Individual decision making has no parallel to organisms being simultaneously subject to different erratically determined probability distributions. Moreover, though rules for permissible actions look like rules

for determining fittest traits, one always must be able to choose some action,[10] while there need not be any fittest traits. Finally, while there are clear standards for evolutionary success in some contexts, part of what is under contention in imprecise decision theory is the standard by which to evaluate rules for choosing gambles—and perhaps what rationality itself consists in.

The work presented in this paper is, I hope, a starting point for further developments. Among other things, my arguments should be generalized for fitnesses that can't easily be represented as expectations.

References

Abrams, Marshall (2009a). The unity of fitness. *Philosophy of Science* 76(5):750–761.

Abrams, Marshall (2009b). What determines biological fitness? The problem of the reference environment. *Synthese* 166(1):21–40.

Abrams, Marshall (2012). Mechanistic probability. *Synthese* 187(2):343–375.

Abrams, Marshall (2014). Environmental grain, organism fitness, and type fitness. In Trevor Pearce; Gillian A. Barker; and Eric Desjardins, eds., *Entangled Life: Organism and Environment in the Biological and Social Sciences*, History, Philosophy and Theory of the Life Sciences. Springer.

Abrams, Marshall (2015). Probability and manipulation: Evolution and simulation in applied population genetics. *Erkenntnis* 80(S3):519–549.

Alexander, J. McKenzie (2007). *The Structural Evolution of Morality*. Cambridge University Press, Cambridge, UK.

[10] See (Joyce, 2010; Bradley, 2015a), but cf. (Rinard, 2015).

Augustin, Thomas; Coolen, Frank P. A.; Cooman, Gert de; and Troffaes, Matthias C. M., eds. (2014). *Introduction to Imprecise Probabilities*. Wiley.

Beatty, John and Finsen, Susan (1989). Rethinking the propensity interpretation: A peek inside Pandora's box. In Michael Ruse, ed., *What the Philosophy of Biology is*, pp. 17–30. Kluwer Academic Publishers.

Beisbart, Claus (online January 2016). A Humean guide to Spielraum probabilities. *Journal of General Philosophy of Science* .

Bradley, Seamus (2015a). How to choose among choice functions. In *ISIPTA '15: Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications*, pp. 57–66.
URL http://www.sipta.org/isipta15/data/paper/9.pdf

Bradley, Seamus (2015b). Imprecise probabilities. In Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy*. Summer 2015 ed.

Bradley, Seamus and Steele, Katie (2016). Can free evidence be bad? Value of information for the imprecise probabilist. *Philosophy of Science* 83(1):1–28.

Brandon, Robert N. (1990). *Adaptation and Environment*. Princeton University, Princeton, New Jersey.

Dardashti, Radin; Glynn, Luke; Thebault, Karim; and Frisch, Mathias (2014). Unsharp Humean chances in statistical physics: A reply to Beisbart. In Maria Carla Galavotti; Dennis Dieks; Wenceslao J. Gonzalez; Stephan Hartmann; Thomas Uebel; and Marcel Weber, eds., *New Directions in the Philosophy of Science*, pp. 531–542. Sprinter.

Fierens, Pablo I. (2009). An extension of chaotic probability models to real-valued variables. *International Journal of Approximate Reasoning* 50:627–641.

Fierens, Pablo I.; Rêgo, Leandro Chaves; and Fine, Terrence L. (2009). A frequentist understand of sets of measures. *Journal of Statistical Planning and Inference* 139:1879–1892.

Giacomini, Henrique C. and Shuter, Brian J. (2013). Adaptive responses of energy storage and fish life histories to climatic gradients. *Journal of Theoretical Biology* 339:100–111.

Gillespie, John H. (1977). Natural selection for variances in offspring numbers: A new evolutionary principle. *American Naturalist* 111:1010–1014.

Gillespie, John H. (1998). *Population Genetics: A Concise Guide*. The Johns Hopkins University Press.

Godfrey-Smith, Peter (2009). *Darwinian Populations and Natural Selection*. Oxford University Press, Oxford, UK.

Hájek, Alan and Smithson, Michael (2012). Rationality and indeterminate probabilities. *Synthese* 187(1):33–48.

Huntley, Nathan; Hable, Robert; and Troffaes, Matthias C. M. (2014). Decision making. In Augustin et al. (2014), chap. 8, pp. 190–206.

Joyce, James M. (2010). A defense of imprecise credences in inference and decision making. *Philosophical Perspectives* 24(1):281–323.

Levi, Isaac (1980). *The Enterprise of Knowledge*. MIT Press, Cambridge, Massachusetts.

Levins, Richard (1968). *Evolution in Changing Environments*. Princeton.

Lewis, David (1994). Humean supervenience debugged. *Mind* 103:473–90.

Lewontin, Richard C. (1970). The units of selection. *Annual Review of Ecology and Systematics* 1:1–18.

Li, Ming and Vitányi, Paul (2008). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, 3rd ed.

Mills, Susan and Beatty, John (1979). The propensity interpretation of fitness. *Philosophy of Science* 46(2):263–286.

Moss, Sarah (2015). Credal dilemmas. *Noûs* 49(4):665–683.

Okasha, Samir (2007). Rational choice, risk aversion, and evolution. *The Journal of Philosophy* 104(5):217–235.

Okasha, Samir and Binmore, Ken, eds. (2012). *Evolution and Rationality: Decisions, Cooperation and Strategic Behavior*. Cambridge University Press.

Papamarcou, Adrianos and Fine, Terrence L. (1986). A note on undominated lower probabilities. *The Annals of Probability* 14(2):710–723.

Pence, Charles H. and Ramsey, Grant (2013). A new foundation for the propensity interpretation of fitness. *The British Journal for the Philosophy of Science* .

Ramsey, Grant (2006). Block fitness. *Studies in History and Philosophy of Biological and Biomedical Sciences* 37(3):484–498.

Rinard, Susanna (2015). A decision theory for imprecise probabilities. *The Philosophers Imprint* 15(7):1–16.

Rosenthal, Jacob (2012). Probabilities as ratios of ranges in initial-state spaces. *Journal of Logic, Language, and Inference* 21:217–236.

Schoenfield, Miriam (2012). Chilling out on epistemic rationality. *Philosophical Studies* 158(2):197–219.

Sober, Elliott (2001). The two faces of fitness. In Rama S. Singh; Costas B. Krimbas; Diane B. Paul; and John Beatty, eds., *Thinking About Evolution*, pp. 309–321. Cambridge University Press, Cambridge, UK.

Stearns, Stephen C. (1992). *The Evolution of Life Histories*. Oxford University Press.

Strevens, Michael (2011). Probability out of determinism. In Claus Beisbart and Stephann Hartmann, eds., *Probabilities in Physics*, chap. 13, pp. 339–364. Oxford University Press, Oxford, UK.

Troffaes, Matthias C. M. (2007). Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning* 45:17–29.

Troffaes, Matthias C. M. and de Cooman, Gert (2014). *Lower Previsions*. Wiley.

Wagner, Günter P. (2010). The measurement theory of fitness. *Evolution* 64(5):1358–1376.

Walley, Peter (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall.

Walley, Peter and Fine, Terrence L. (1982). Towards a frequentist theory of upper and lower probability. *The Annals of Statistics* 10(3):741–761.

Walsh, Denis M. (2013). Descriptions and models: Some responses to Abrams. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 44(3):302–308.

White, Roger (2009). Evidential symmetry and mushy credence. In Tamar Szabó Gender and John Hawthorne, eds., *Oxford Studies in Epistemology*, vol. 3, pp. 161–186. Oxford University Press.

Scientific Realism and Primitive Ontology

Valia Allori
Northern Illinois University

Abstract

In this paper I wish to connect the recent debate in the philosophy of quantum mechanics concerning the nature of the wave-function to the historical debate in the philosophy of science regarding the tenability of scientific realism. Being realist about quantum mechanics is particularly challenging when focusing on the wave-function. According to the wave-function ontology approach, the wave-function is a concrete physical entity. In contrast, according to an alternative viewpoint, namely the primitive ontology approach, the wave-function does not represent physical entities. In this paper, I argue that the primitive ontology approach can naturally be interpreted as an instance of the so-called 'explanationism' realism, which has been proposed as a response to the pessimistic-meta induction argument against scientific realism. If my arguments are sound, then one could conclude that: (1) contrarily to what is commonly though, if explanationism realism is a good response to the pessimistic-meta induction argument, it can be straightforwardly extended also to the quantum domain; (2) the primitive ontology approach is in better shape than the wave-function ontology approach in resisting the pessimistic-meta induction argument against scientific realism.

1. Introduction

Scientific realism would be a commonsensical philosophical position if there weren't powerful counter-arguments to it, the most famous of which is the pessimistic meta-induction (PMI) argument: since past successful theories turned out to be false, it is unwarranted to believe that our current theories are true simply because they are successful [Laudan 1981]. Some scientific realists have responded to the PMI argument by restricting realism to a subset of the theoretical entities of the theory. One particular way of doing this is explanationism realism (ER), according to which one should be realist with respect to the working posits of the theory, namely the ones involved in explanations and predictions and that are preserved in theory change. In contrast, one does not need to commit herself to believe in the existence of other presuppositional posits that theory makes, since they are somewhat 'idle' components [Kitcher 1993], [Psillos 1999].

The proponents of this version of restricted (or localized, or selective, or preservative) realism focus on examples like Fresnel's theory of light that postulated the existence of ether, and argue that ether hasn't played a crucial role in the success of the theory, and did not carry over to Maxwell's electrodynamics. Because of this, the realist should commit to the existence of electromagnetic waves but not to the existence of ether. Similar arguments have been put forward for phlogiston and caloric. I think that analyzing these 'classical' examples is important and interesting; nonetheless, the case

for ER is fundamentally incomplete if one does not show that also in the theory change from classical to quantum mechanics the working posits of classical mechanics are preserved in quantum theory, and that they play an essential role in the predictions and explanations of both theories. In this paper, I argue that ER can be extended to the quantum framework. In order to show this, I discuss the different realist approaches to quantum mechanics, which fundamentally differ in the interpretation they provide of (what seems to be) the fundamental object of quantum mechanics, namely the wave-function. On the one hand, according to the so-called wave-function ontology (WFO) approach, the wave-function is a concrete physical entity [Albert 1996], [Ney 2013], [Lewis 2004]. In contrast, according to the primitive ontology (PO) approach [Allori et al. 2008], [Allori 2013], the wave-function does not represent physical objects. I argue that the PO approach can naturally provide what ER needs to defeat the PMI argument when applied to the transition from classical to quantum physics. In fact the PO can be identified with the working posits of the theories, and as such: (1) it is primarily responsible for the success of both classical and quantum mechanics, and (2) the PO is (suitably) preserved in the classical-to-quantum transition. Therefore, the realist should commit to the existence of the PO, while she can be 'neutral' with respect to the other theoretical components of both theories. In this way, the PO approach provides an interesting framework for the scientific realist, given that it allows ER to naturally extend to the quantum domain.

To conclude, I wish to underline that ER so understood provides an argument in favor of the PO approach when compared to the WFO approach: in virtue of being preserved in theory change and playing a crucial role in the success of the old and the new theory, the PO does not fall prey of the PMI argument. In contrast, if one insists, like a proponent to the WFO approach would, that the wave-function represents physical objects, then it is hard to see how the working posits can be the same in both theories, given that the wave-function does not have any classical analog. Because of this, the wave-function ontologist seems to be in trouble: if the ontology of quantum mechanics is fundamentally different from the one of quantum theory, how can we respond to the PMI argument? Other options could be available to the proponent of the WFO approach, like for instance structural realism, but surely not ER, which is available only to the primitive ontologist.

The paper is structured as follows: in the next section, there is an overview of the PMI argument and of the replies to it in terms of restriction of realism, with a particular emphasis to ER. Then in Section 3, the discussion focuses on the necessity of extending ER to the quantum domain. The PO approach is presented and succinctly discussed, in Section 4, underlining how the PO is preserved through theory change and how it is fundamentally responsible for the empirical and theoretical success of the theory. The last section discusses the advantage of the PO approach over the WFO approach in responding to the PMI argument: while the primitive ontologist can use an

explanationist realist strategy, the wave-function ontologist will have to find something else.

2. The Pessimistic Meta Induction and Explanationism Realism

Scientific realism is, roughly put, the view that scientific theories give us a (nearly) truthful description of the world. So, scientific theories discuss the behavior of a number of unobservable entities (e.g. electrons), and the scientific realist claims that we have good reasons to consider these entities as truly existing. The main argument for scientific realism, the no-miracle argument (NMA), can be summarized as follows: "realism is the only philosophy that does not make the success of science a miracle" [Putnam 1975: 73]. That is, the empirical success of a theory can and should be taken as evidence of its truth. Nonetheless, there are very powerful arguments against scientific realism, one of them being Laudan's PMI argument. The main idea is that it is not the case that, against the NMA, the empirical success of a theory is a reliable indicator of its truth. Here is a way to spell the argument out, as a proof by contradiction:

Premise 1: Empirical success is a reliable indicator of truth (*reductio* assumption);

Premise 2: Our most current theory is true;

Premise 3: Most past successful theories are false;

Conclusion: Therefore, empirical success is not an indicator of truth.

More succinctly: our current theories, even if successful, are more likely to be false than true since many past theories were successful but false.

One way to respond to the PMI challenge is to restrict realism, and argue that one should be realist about a restricted set of entities, not about the whole theory. This is what Psillos calls a *divide et impera* strategy: scientific realists may argue that "when a theory is abandoned, its theoretical constituents, i.e. the theoretical mechanisms and laws posited, should not be rejected *en bloc*. Some of those theoretical constituents are inconsistent with what we now accept, and therefore they have to be rejected. But not all are. Some of them have been retained as essential constituents of subsequent theories" [Psillos 1991: 108]. So, if one can show that the entities that are retained in moving from one theory to the next are the ones that are responsible for the empirical success of the theory, the PMI argument is blocked. In fact, this argument works only if one assumes that past theories are false in their entirety, even if they were successful, so that their success has to come from something else other than their truth. By restricting realism, instead, one provides an alternative explanation for the success of past false successful theories: past theories were successful not because they were (approximately) true in their entirety, but because some parts of them were. And these true constituents of past theories are responsible for the theories' success and they are carried over in theory change. Because of this, we are justified in believing that the entities these theoretical constituents represent really exist.

There are various ways to restrict realism, namely, there are different ways to resist the PMI by limiting the number or the kind of entities in the theory the realist should be

taking 'ontologically seriously.' One such example is Worrall's structural realism (SR) [Worrall 1989]. According to this view, what is preserved in theory change is the mathematical structure of the theory, rather than the theoretical content of the theory: the PMI is correct in saying that we often have discontinuity in theory change at the level of unobservable entities, but most of the mathematical content of the old theory carries over to the new one. Because of this, the scientific realist may not be justified in believing what the theory says about the nature of physical objects, nonetheless she is justified in believing that the structure that holds between these objects which is preserved in theory change is (approximately) true. There are different varieties of SR, but a first rough distinction is the one between epistemic SR and ontic SR. In the epistemic version, which some attribute to Worrall himself, the claim is that we do not have justification for believing that objects have the nature our theories suggest they have, but we are only justified in believing that these objects stand in certain structural relations with one another. Ontic SR instead goes further and claims that the very notion of objects is problematic and is worth dismissing [French 1998], [Ladyman 1998].

There are other responses to the PMI argument[1], but in this paper I will focus on ER, developed most prominently by [Psillos 1999] and [Kitcher 1993]. They distinguish between 'working' and 'presuppositional' posits (or 'idle wheels') of a theory, and argue that one should be realist about the working posits but not the presuppositional posits, and this is because these posits are not involved in the success of the theory. In fact, if one analyzes the mechanism of empirical success of past theories one will see, they argue, that only certain entities are involved, namely the working posits. The theory postulates the existence of other entities too, for a variety of reasons, but these entities are never used to derive predictions or to provide explanations in the framework of the theory. If the working posits are preserved during theory change, while the presuppositional posits are not, the argument goes, one is justified in believing in the working posits of a theory exist. The other theoretical constituents, the presuppositional posits, are 'idle' components, which make no difference to the theory's success and thus the realist has no need to commit to.

In this way one can resist to the PMI argument: past theories were successful because they got something right, namely the working posits, but they are also false when considered in their entirety because they got something wrong too, namely the presuppositional posits.

[1] Most notably, another restriction of realism is entity realism [Hacking 1982], according to which realist commitments should be limited to unobservable entities that could be causally manipulated. In addition semirealism [Chakravartty 2007], which in certain respects is in-between structuralism and entity realism, recommends realism with respect of the so-called detection properties, namely the properties in the theory which are tied to our perceptual and causal experiences, and not to auxiliary properties, which are not essential in establishing existence claims. Therefore, one should restrict realism to the detection properties.

3. The Classical-to-Quantum Theory Change as a Problem for Explanationsim Realism

Scientific realism has been motivated and discussed almost exclusively discussing theories other than quantum mechanics. In particular, Psillos and Kitcher argue for ER discussing that Fresnel's theory of light was successful because it got the working posit right, namely the electromagnetic waves: they are responsible for the success of the theory, and they were preserved by Maxwell's electrodynamics. In contrast, ether was a presuppositional posit: the success of Fresnel's theory did not depend on it, and it was abandoned by the subsequent theory. Realists are therefore justified in believing that electromagnetic waves exist, but do not have to be committed to believe that ether exists too. Another example extensively discussed in the literature is the caloric theory of heat, or phlogiston's theory of combustion, to again arrive to the conclusion that caloric and phlogiston are presuppositional posits. In reply, these historical examples have been revisited with the intent of arguing that ether, caloric, phlogiston, and the like, contrarily to what it is maintained by ER, played an important role in the success of past theories (see, e.g. [Elsamahi 2005], [Chang 2003]).

Regardless of what the outcome of the debate over these examples is, it seems to me that the main threat to ER comes from the transition from classical to quantum theories. The fact that the discussion was limited to 'classical' theories is, in a sense, not surprising: quantum mechanics has been considered, for a long time, incompatible with realism: while, on the one hand, quantum theory is incredibly powerful in making new and very precise predictions, on the other hand it is extremely difficult to understand what image of the world it provides us. Indeed, quantum mechanics has been taken by many to suggest that physical objects have contradictory properties, like being in a place and not being in that place at the same time, or that properties do not exist at all independently of observation. Given that, many have thought that the real lesson of quantum mechanics is that the dream of the scientific realist is impossible: the theory is extremely successful, but it seems impossible to explain this success in terms of the theory being (approximately) true, unless one is willing to give up, say, classical logic or the like to account for the existence of objects with contradictory properties. Luckily, the situation has changed: today we have various proposal of quantum theories that allow for a realist reading. Among these theories, most famously we find Bohmian and Everettian mechanics (BM and EM respectively), and the GRW theory (GRW): they are empirically adequate fundamental physical theories according to which there is an objective physical world, which can be described by non-contradictory, mind-independent properties.

The problem for ER is that even assuming that one could be a realist with respect to quantum mechanics, quite strikingly, when examples from quantum mechanics are discussed, they are brought up to motivate ontic SR rather than ER: "we have learned from contemporary physics is that the nature of space, time and matter are not

compatible with standard metaphysical views about the ontological relationship between individuals, intrinsic properties and relations" [Ladyman 2014]. In addition, and presumably more importantly, it does not seem obviously the case, as an explanationist realist would have to maintain, that some theoretical entities of past theories are carried over to quantum mechanics, and they are the ones responsible for quantum mechanics' enormous success. Indeed, exactly the opposite seems to be true: in quantum mechanics we have the Schrödinger equation, which is the evolution equation of the wave-function, an object which is involved in the derivation of most, if not all, predictions and explanations the theory is able to provide, and which arguably does not have any classical analog. If so, ER seems to be doomed: not only the wave-function is something new to classical mechanics, but it seems to be the fundamental object that drives quantum mechanics in all its explanations and predictions. We have radical discontinuity and therefore the PMI argument has not been blocked.

In light of all this, I think that case for ER has no hope of being compelling if does not cover quantum mechanics. In the next section I show how ER can be extended to quantum theories if paired with a particular view about the metaphysics of quantum mechanics, namely the PO approach.

4. Primitive Ontology and Explanationism Realism

Most philosophers of physics recognize the legitimacy of BM, EM and GRW, but disagree about the metaphysical pictures these theories actually provide. In this section, I wish to show how the PO theories provide examples of quantum theories with the same (or suitably similar) working posits as classical mechanics. That is, the claims are going to be that: (1) the PO is the primary responsible of the theory's success; and (2) the PO (suitably) carries over during theory change. If so, assuming that a strategy like ER is successful in defending scientific realism, the PMI argument is blocked: the realist is justified in believing that the PO is real because it does all the work to explain empirical success of theories and it is preserved in theory change.

Here is a brief summary of the PO account [Allori et al. 2008]. In quantum theories understood within the PO framework, there are two fundamental ingredients that are supposed to represent, respectively, what matter is, and how matter behaves. Matter is represented by entities in three-dimensional space (or four-dimensional space-time), which are the PO of the theory. Examples of possible primitive ontologies include point-particles, continuous fields, and spatio-temporal events (flashes). Quantum theories with different primitive ontologies are discussed and analyzed by the proponents of the PO approach in different papers, and some examples are worth mentioning: BM is a theory with a particle PO, GRWm is a theory in which matter is described by a continuous (three-dimensional) matter field localized where the macroscopic objects are, while GRWf is a theory of flashes, namely discrete spatio-temporal events. How matter behaves is explicated in terms of the law of evolution of the PO, which in turn is implemented by the so-called 'nomological' variables, most

importantly by the wave function. Therefore, even if the wave-function evolves in time (according to either the Schrödinger equation or some variant of it), it *never represent matter*. One cannot dispense of the wave function from quantum theories[2] but that does not mean, according to the proponents of this approach, that we should think the wave-function represents material objects. Rather, it is a necessary ingredient to implement the law of temporal evolution of the PO [Allori 2013]. To continue with the examples mentioned above, we have that in BM the wave-function evolves according to the Schrödinger equation, while in GRWf and GRWm it evolves according to Schrödinger equation and then randomly collapses, following the so-called GRW evolution.
Here are some fundamental features of the PO approach that is crucial to articulate:

(1) [REDUCTIONISM wrt PO] The motivation of the PO approach is to account for the existence of macroscopic objects, which are thought to be fundamentally composed of the microscopic entities the PO specifies. As such, the PO approach is reductionist, at least to the extent that it allows to make sense of claims like the PO being "the building blocks of everything else," and of the idea that macroscopic regularities are obtained entirely from the microscopic trajectories of the PO.

(2) [EXPLANATION and PO] The PO explains the macroscopic regularities using reductionist approaches similar to those used in classical mechanics. In fact, in classical mechanics, macroscopic bodies are made of a collection of particles, and their properties are accounted for in terms of the interaction of these particles among each other and the particles of the environment. For instance, the transparency of a pair of glasses is explained in terms of the electromagnetic forces acting between the particles composing the glasses, which are such that incoming light rays will pass through them. Similarly, the PO grounds the explanatory schema of quantum theories: macroscopic objects are made of entities described by the PO, and their properties are in principle accounted for in terms of the PO's behavior (see [Allori 2013]).

(3) [THEOREITICAL VIRTUES] What variable is the PO of a theory is postulated, rather than inferred from the formalism. One PO is chosen over another on the basis of some super-empirical virtues such as simplicity, explanatory power, and unification: the PO that provides the simplest, most unifying explanation should be selected (see [Allori 2015]). Because macroscopic regularities are accounted for in terms of PO and because the role of the wave-function is to implement the law for the PO, the nature of the PO (particles, field, flashes,…) is not necessarily connected to the law of evolution of the wave-function (Schrödinger, GRW…): for instance in BM the PO of particles is connected with

[2] To be precise, some attempts have been made to eliminate the wave-function entirely from quantum theories (see e.g. [Dowker Herbauts 2005] and [Norsen 2010]).

a Schrödinger evolving wave-function, but one can imagine a theory of particles with a GRW-evolving wave-function, (see [Allori et al 2014] for more examples).

(4) [UNDERDETERMINATION of PSI] The way the wave-function evolves in time is irrelevant as long as the law of the PO such wave-function defines remains the same: a theory of particles which follow certain trajectories, like BM, can be obtained by a Schrödinger-evolving wave-function, like in the usual formulation, but also in terms of a collapsed wave-function (see [Allori et al 2008] for details). Two theories with the same trajectories for the PO, regardless of how they have been obtained (i.e., via a Schrödinger evolving wave-function or not) are called physically equivalent. Since different wave-functions can give rise to the same trajectories for the PO, and since the trajectories of the PO are the ones that account for the macroscopic regularities, the wave-function evolution is underdetermined by the data.

(5) [PREDICTIONS] Once the PO and its law of evolution have been chosen, everything else is determined, including the empirical predictions which are determined as a function of the PO. The wave-function appears into the derivation of the predictions of the theory, but its role is not essential, since as we just saw, the way in which it specifies the law of the PO is underdetermined(see [Allori et al 2014] for more on this point).

Now, the idea that I wish to put forward is that there are striking similarities between the PO approach and ER. In particular, it seems to me that the PO can be identified with the working posit of quantum mechanics, while the wave-function is best seen as a presuppisitional posit. In fact, as we saw in (5) above, the predictions are determined by the PO, not by the wave-function, which does appear in the derivation, but whose evolution is underdetermined by data, as we saw in (4). In addition, as we saw in (2), explanation is in terms of the PO, and this reminds of Kitcher's idea that working posits are the entities that play a fundamental role in the theory's explanatory schemata. Moreover, there is the explicit fundamental postulation that the PO represent matter, while the wave-function does not, and that everything is made of the entities the PO specifies, as we saw in (1). The PO approach suggests we should be realist about the PO, regardless of what we think the wave-function really is. In fact, all primitive ontologists (or supporters of suitably related views) maintain that one should be realist about the PO, but they have different ideas about the wave-function: it has been considered to be, among other things, a law-like object [Goldstein et al. 2013], a disposition [Esfeld et al. 2014], a property [Monton 2006], or a new kind of entity [Maudlin 2013]. Nonetheless, one can be 'metaphysically neutral' with respect to the wave function: one does not need to postulate the existence of the wave-function in order to account for the success of the theory. But this is to say that the PO is a working posit, while the wave-function is a presuppositional posit of quantum theories. If so, the PO approach provides a very

nice framework for the explanationist realist to extent her view in the quantum domain: one should be realist about the PO because it is the sole responsible for the theory's success.

However, this is not enough to successfully reply to the PMI argument: one would have to show that the PO is preserved during theory change. What is the PO of classical mechanics? The answer seems to be straightforward: according to classical mechanics, matter is made of particles, intended as objects with the fundamental property of having a position in three-dimensional space. Therefore, since we do not need to worry about the wave function, the preservation of PO during the classical-to-quantum theory change is obvious for quantum theories of particles, like BM. The interesting and more challenging cases are the ones that involve POs different than that, namely a mater density PO or a flash PO. In both cases, literally, the PO of classical mechanics has not carried over. The nature of the objects the theories postulate is fundamentally different: on the one hand in classical mechanics we have particles fundamentally identified by having a definite position in space and following given trajectories in space-time; on the other hand, we have either a continuous matter filed in GRWm, or a discrete set of spatio-temporal event in GRWf. Should we think this is an instance of radical discontinuity, and should we take this to be a reason to give up on ER? I believe this would be too harsh: what seems to be true is not necessarily that there are particles, or fields or flashes, but rather that there is 'stuff' in three-dimensional space, and this is what matter is. When there was the theory change from the theory that atoms are indivisible to the theory that atoms are made of other particles which themselves are thought as indivisible, one could maintain that what the theory got right is that there are particles, but it was wrong about which the fundamental particles really were: we thought they were atoms, but they are neutron, protons and electrons instead. The situation here is slightly different, being more similar to the case in which we move to a theory in which the fundamental entities are particles, to a theory in which the fundamental entities are instead strings. What are we getting right here? Not the nature of the fundamental: before we had one-dimensional particles, now we have bi-dimensional vibrating loops. However, I think it is important to underline that if we 'squint,' then we don't see the fine-grained details, and we take strings to be particles. They are, for all *explanatory* purposes, particles: we need to explain the macroscopic regularities, and we explain them in terms of the PO ignoring the details about what composes it. Just like when we observe a hose from a distance and we think it is one-dimensional while it is actually two-dimensional, or when we look at a poster in the subway and we think it's an image while instead it is a collection of colored dots. At the level of microphysics we may have flashes or a continuous field, but at some mesoscopic level they produce trajectories as if they are produced by particles. So, even if the microscopic PO is not one of particles, there is a mesoscopic scale in which they

behave as if they are, and then from that level up to the macroscopic level, the explanation is the same as if they were particles.

The obvious worry here is: isn't that just some sort of (ontic) structuralism? If we do not preserve the nature of 'stuff,' isn't what we preserve some structural content of the theory? If structuralism is the view that there is just structure and no objects, then clearly not, since the PO approach postulates the existence of objects as a starting point. What about a moderate version of ontic SR, like the one proposed by [Esfeld 2004]? The idea behind this view is something like this: one should be realist about structure but, in contrast with the 'eliminativist' ontic SR mentioned above, there are 'things' that stand in the relation the structure prescribes, even if they have no intrinsic identities. In the quantum domain, such structure is the wave-function. Indeed, interestingly enough, [Esfeld forthcoming] proposes that in his moderate ontic SR, the *relata* the wave-function relates are given by the PO: he argues that the PO approach and his moderate ontic SR can help each other make sense of quantum non-locality and entanglement. So, in his view, we should be realist about the PO, and also about the structure that relates the PO, provided by the wave-function. In this sense, the reading I provide of the PO approach is not structural: the strength of the PO approach in responding to the PMI argument is that it regards the wave-function as a working posit. Only because of this, one can show there is continuity of PO during theory change. Instead Esfeld's moderate ontic SR does not have this advantage: if the wave function is the structure the realist should be committing to, then it is difficult to see where this structure was coming from in classical physics.

The PO approach entails something like this: we do not get the nature of objects right because we believe they are particles in classical mechanics and then we discover they are actually, say, flashes in quantum mechanics; but we get some 'structure' right, namely that on some mescoscopic level they behave as if their nature were the one of particles. One may call it structural realism, but it does not seem to have much in common with the other varieties of SR we just examined. Rather, more appropriate in my opinion is the connection with ER: what provides the explanation, namely the PO, is what 'ontologically counts,' if it is preserved in theory change.

Another interesting question is whether the PO approach can help reply to some of the objections that have been raised to ER, most notably the problem that it is not clear whether it is possible to precisely and objectively identify the working posits of a theory rather than doing that post hoc: the working posits are what we see have carried over (see, e.g. [Stanford 2003a, 2003b]). Indeed, the PO approach seems to provide an improvement with respect to ER: the PO is postulated when the theory is proposed, rather than inferred from the formalism, as the one that provides the best combination of simplicity, explanatory and unificatory account of the experimental data. In this way, what is a working posit is selected from the time the theory is proposed, and it is never

"post hoc." If the PO, together with the explanatory schema, is preserved during theory change, then there is no radical discontinuity and the PO is truly a working posits.

5. A New Argument for the PO Approach

To summarize the last section, I have shown how the PO approach may naturally be seen as an instance of ER in which one restricts realism to the PO: since the PO is carried over in theory change, and it is the primary responsible for the theory empirical success, then one is justified in believing the entities it represents really exists. As such, the PO approach provides the ER with a straightforward strategy to block the PMI in the quantum framework.

In this section, I wish to notice that this analysis of the PO approach as an instance of ER also provides the PO approach with an important advantage over the alternative WFO approach. According to this view, the wave-function is a concrete physical field and should be regarded as representing matter. If we analyze this view in terms of ER, therefore, the wave-function has to be a working posit of quantum theory. The problem with this is that, mathematically, the wave-function is an object that lives in the highly dimensional configuration space, and as such is a very different entity from classical particles. In addition, the image of the world provided by the WFO approach is very different from the image of the world given to us by classical mechanics: in the latter there are particles moving in three-dimensional space, in the former there is this matter field in a highly-dimensional space. In the classical-to-quantum transition we discover that not only we were getting the nature of objects wrong (we believed there were particles and actually there are none) but we cannot get our classical picture back by 'squinting,' like in the PO framework, since the fundamental physical space is not three-dimensional. In this way, there is no continuity of working posits between classical and quantum mechanics, and the strategy to resist to the PMI argument along the lines of ER is precluded to the proponent of the WFO approach. If there is truly a quantum revolution, as the WFO approach seems to maintain to a give extent (see [Allori 2015] for an interesting take on this), and the way in which we understand the word using quantum theory is fundamentally different from the way in which we understood it in classical terms, what is our justification to believe that the theoretical terms used in quantum mechanics are (approximately) true? It is difficult to see how the PMI could be defeated in the WFO framework, unless they go eliminative structural realists, and they may not want to do that, given the numerous objection that have been raised against this view (see, e.g. [Psillos 1999], [Chakravartty 2007], [Cao 2003] among others).

Word count: 5660

References

- [Albert 1996] Albert, David Z. "Elementary Quantum Metaphysics." In: J. Cushing, A. Fine and S. Goldstein (eds.), *Bohmian Mechanics and Quantum Theory: An Appraisal*. Boston Studies in the Philosophy of Science 184: 277-284 (1996).
- [Albert Ney 2013] Albert, David Z., Ney, Alyssa (eds.). *The Wave Function.* Oxford University Press (2013).
- [Allori et al. 2008] Allori, Valia, Sheldon Goldstein, Roderich Tumulka, and Nino Zanghi. "On the Common Structure of Bohmian Mechanics and the Ghirardi-Rimini-Weber Theory." *The British Journal for the Philosophy of Science* 59 (3): 353-389 (2008).
- [Allori et al. 2014] Allori, Valia, Sheldon Goldstein, Roderich Tumulka, and Nino Zanghi. "Predictions and Primitive Ontology in Quantum Foundations: A Study of Examples." *The British Journal for the Philosophy of Science* 65 (2): 323-352 (2014).
- [Allori 2013] Allori, Valia. "Primitive Ontology and the Structure of Fundamental Physical Theories." In: D. Z. Albert, A. Ney (eds.), *The Wave Function.* Oxford University Press: 58-75 (2013).
- [Allori 2015] Allori, Valia. "Quantum Mechanics and Paradigm Shifts." *Topoi* 32 (2): 313-323 (2015).
- [Cao 2003] Cao, Tian Yu. "Can We Dissolve Physical Entities into Mathematical Structures?" In: J. Symonds (ed.), Special Issue: Structural Realism and Quantum Field Theory. *Synthese* 136 (1): 57–71 (2003).
- [Chakravartty 2007] Chakravartty, Anjan. *A Metaphysics for Scientific Realism: Knowing the Unobservable*. Cambridge University Press (2007).
- [Chang 2003] Chang, Hasok. "Preservative Realism and Its Discontents: Revisiting Caloric." *Philosophy of Science* 70: 902-912 (2003).
- [Dowker Herbauts 2005] Dowker, Fay, Isabelle Herbauts. "The Status of the Wave Function in Dynamical Collapse Models." *Foundations of Physics Letters* 18: 499–518 (2005).
- [Elsamahi 2005] Elsamahi, Mohamed. "A Critique of Localized Realism." *Philosophy of Science* 72 (5): 1350-1360 (2005).
- [Esfeld forthcoming] Esfeld Michael. "How to Account for Quantum Non-locality: Ontic Structural Realism and the Primitive Ontology of Quantum Physics." *Synthese* (forthcoming).
- [Esfeld et al. 2014] Esfeld, Michael, Dustin Lazarovici, Mario Hubert, and Detlef Dürr: "The Ontology of Bohmian Mechanics." *The British Journal for the Philosophy of Science* 65: 773–796 (2014).
- [French 1998] French, Stephen. "On the Withering Away of Physical Objects." In" E. Castellani (ed.), *Interpreting Bodies: Classical and Quantum Objects in Modern Physics*. Princeton University Press: 93–113 (1998).

- [Goldstein et al. 2013] Goldstein, Sheldon, and Nino Zanghi. "Reality and the Role of the Wave Function in Quantum Theory." In: D. Z. Albert, A. Ney (eds.), *The Wave Function.* Oxford University Press: 96-109 (2013).
- [Hacking 1982] Hacking, Ian. "Experimentation and Scientific Realism." *Philosophical Topics* 13: 71–87 (1982).
- [Kitcher 1993] Kitcher, Philip. *The Advancement of Science*. Oxford University Press (1993).
- [Ladyman 1998] Ladyman, James. "What is Structural Realism?" *Studies in History and Philosophy of Science* 29: 409-424 (1998).
- [Ladyman 2014] Ladyman, James. "Structural Realism." The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.). Spring 2014. URL = <http://plato.stanford.edu/archives/spr2014/entries/structural-realism/>.
- [Laudan 1981] Laudan, Larry. "A Confutation of Convergent Realism." *Philosophy of Science* 48 (1): 19-49 (1981).
- [Lewis 2004] Lewis, Peter. "Life in Configuration Space." *The British Journal for the Philosophy of Science* (55): 713–729 (2004).
- [Maudlin 2013] Maudlin, Tim. "The Nature of the Quantum State." In: D.Z. Albert, A. Ney (eds.), *The Wave Function*. Oxford University Press: 126-154 (2013).
- [Monton 2006] Monton, Bradley. "Quantum Mechanics and 3-N Dimensional Space." Philosophy of Science 75 (5): 778-789 (2006).
- [Ney 2012] Ney, Alyssa. "The Status of Our Ordinary Three Dimensions in a Quantum Universe." *Nous* 46 (3): 525–560 (2012).
- [Norsen 2010] Norsen, Travis. "A Theory of (Exclusively) Local Beables. *Foundations of Physics* 40 (12):1858-1884 (2010).
- [Psillos 1999] Psillos, Stathis. *Scientific Realism: How Science Tracks Truth*. Routledge (1999).
- [Putnam 1975] Putnam, Hilary. *Mathematics, Matter and Method*. Cambridge University Press (1975).
- [Stanford 2003a] Stanford, Kyle. "Pyrrhic Victories for Scientific Realism." *Journal of Philosophy* 100: 553–572 (2003).
- [Stanford 2003b] Stanford, Kyle. "No Refuge for Realism: Selective Confirmation and the History of Science." *Philosophy of Science* 70: 913–925 (2003).
- [Worrall 1989] Worrall, John. "Structural Realism: The Best of Both Worlds?" *Dialectica* 43: 99–124 (1989).

Dissolving the missing heritability problem

Pierrick Bourrat & Qiaoying Lu

Abstract: Heritability estimates obtained in genome-wide association studies (GWAS) are much lower than those of traditional quantitative methods. This has been called the "missing heritability problem". By analyzing and comparing these two kinds of methods, we first show that the estimates obtained by traditional methods involve some terms that GWAS do not. Second, the estimates obtained by GWAS do not take into account epigenetic factors transmitted across generations, whilst they are included in the estimates of traditional quantitative methods. Once these two factors are taken into account, we show that the missing heritability problem can be largely dissolved. Finally, we briefly contextualize our analysis within a current discussion on how non-additive factors relate to the heritability estimates in GWAS.

1. Introduction.

One pervasive problem encountered when estimating the heritability of quantitative traits is that the estimates obtained from Genome-Wide Association Studies (GWAS) are much smaller than that calculated by traditional quantitative methods. This problem has been called the missing heritability problem (Turkheimer 2011). Take human height for example. Traditional quantitative methods deliver a heritability estimate of about 0.8, while the first estimates using GWAS were 0.05 (Maher 2008). More recent GWAS methods have revised this number and estimate the heritability of height to be at most 0.45 (Yang et al. 2010; Turkheimer 2011). Yet, half of the heritability is still missing.

In quantitative genetics, heritability is defined as the portion of phenotypic variation in a population that is caused by genetic difference (Downes 2015). Traditionally, this portion is estimated by measuring the phenotypic resemblance of genetically related individuals without identifying at the molecular level (more particularly the DNA level) the genetic causes of phenotypic variation. GWAS have been developed in order to locate the DNA sequences that influence the target trait and estimate their effects, especially for common complex diseases such as obesity, diabetes and heart disease (Visscher et al. 2012; Frazer et al. 2009). As for height, almost 300 000 common DNA variants in human populations that associate with it have been identified by GWAS (Yang et al. 2010). Granted by many that the heritability estimates obtained

by traditional quantitative methods are quite reliable, the method(s) used in GWAS have been questioned (Eichler et al. 2010).

A number of partial solutions to the missing heritability problem have been proposed, with most of them focusing on improving the methodological aspects of GWAS in order to provide a more accurate estimate (e.g., Manolio et al. 2009; Eichler et al. 2010). Some authors have also suggested that heritable epigenetic factors might account for part of the missing heritability. For instance, in Eichler et al. (2000, 488), Kong notes that "[e]pigenetic effects beyond imprinting that are sequence-independent and that might be environmentally induced but can be transmitted for one or more generations could contribute to missing heritability." Furrow et al. (2011) also claim that "[e]pigenetic variation, inherited both directly and through shared environmental effects, may make a key contribution to the missing heritability." Others have made the same point (e.g., McCarthy and Hirschhorn 2008; Johannes et al. 2008). Yet, in the face of this idea one might notice what appears to be a contradiction: how can *epigenetic* factors account for the missing heritability, if the heritability is about *genes*?

To answer this question as well as to analyze the missing heritability problem, we compare the assumptions underlying both heritability estimates in traditional quantitative methods and those in GWAS. We argue that a) the heritability estimates of traditional methods include some terms associated with broad-sense heritability (H^2), as opposed to narrow-sense heritability (h^2); b) although GWAS are supposed to get h^2, h^2 relies on an evolutionary concept of the gene

that can include epigenetic factors while heritability estimates obtained from GWAS do not. With these two points being illustrated, we expect the missing heritability problem to be largely dissolved as well as setting the stage for further discussions.

The reminder of the paper will be divided into three parts. First, we briefly introduce how heritability is estimated in two traditional methods, namely twin studies and parent-offspring regression. We show that the estimates obtained by each methods include *some* non-additive elements and consequently correspond neither to H^2 nor to h^2, but to a notion in between which we term "broader-sense heritability". Second, we outline the basic rationale underlying GWAS and illustrate that they estimate heritability by considering solely DNA variants. By arguing that the notion of additive genetic variance does not necessarily refer to DNA sequences but can also refer to epigenetic factors in traditional quantitative methods, we show that the notion of heritability estimated in GWAS is more restrictive than that of traditional quantitative methods, and term this notion "DNA-based narrow-sense heritability". Finally, in Section 4, based on the conclusions from Section 2 and Section 3, we claim that the gap between the heritability estimates of traditional quantitative methods and those of GWAS can be explained away in two major ways. One consists in recognizing that if non-additive variance was removed from the estimates obtained via traditional methods, they would be lower. The other consists in recognizing that if epigenetic factors were taken into account by GWAS, the heritability estimates obtained would be higher. We conclude Section 4 by showing how our analysis sheds

some light on a discussion about the role played by non-additive factors in the missing heritability problem. Because human height has been "the poster child" of the missing heritability problem (Turkheimer 2011, 232), we will use this example to illustrate each of our points.

2. Heritability in Traditional Quantitative Methods.

According to quantitative genetics, the phenotypic variance (V_P) of a population can be explained by two components, its genotypic variance (V_G) and its environmental variance (V_E). In the absence of gene-environment interaction and correlation, we thus have:

$$V_P = V_G + V_E \quad (1)$$

From there broad-sense heritability (H^2) is defined as:

$$H^2 = \frac{V_G}{V_P} \quad (2)$$

V_G can further be portioned into the additive genetic variance (V_A), the dominance genetic variance (V_D) and the epistasis genetic variance (V_I). We have:

$$V_P = V_A + V_D + V_I + V_E \quad (3)$$

where V_A is the variance due to hypothetical genes making an equal and additive contribution to the trait studied (e.g., height). V_D is the variance due to interactions between alleles at one locus for diploid organisms, and V_I is the variance due to interactions between alleles from different loci. V_D and V_I together represent the variance due to particular combinations of genes of an organism.

Since genotypes of sexual organisms recombine at each generation via reproduction, dominance and epistasis effects are not transmitted stably across generations, only additive genetic effects are. Therefore, V_A is the variance due to stably transmitted genetic effects. Narrow-sense heritability (h^2) measures to what extent variation in phenotypes is determined by the variation in genes transmitted from parent(s) to offspring (Falconer and Mackay 1996, 123). It is defined as:

$$h^2 = \frac{V_A}{V_P} \quad (4)$$

h^2 is important in breeding studies and is used by evolutionary theorists who are interested in making evolutionary projections of a trait within a population across generations.

To know h^2, both V_A and V_P must be known. V_P, for most quantitative traits (including height), can be directly estimated by measuring individuals. However, there is no direct way to estimate V_A in traditional quantitative methods. The traditional way to estimate it requires two elements. First, one needs a population-level measure of a phenotypic resemblance of family

relative pairs[1]. This measure is obtained by calculating the *covariance* of the phenotypic values for those pairs. The choice of what sort of relatives to use depends on what data is available. The second element is the genetic relation between family pairs. It indicates the percentage of genetic materials the pairs are expected to share. With these two elements, one can estimate how much the genes shared contribute to the phenotypic resemblance. In a large population with different phenotypes, one can then estimate how much the additive genetic difference contributes to phenotypic difference in this population, which estimates h^2.

For simplicity, traditional quantitative methods usually assume that there is neither gene-environment interaction nor correlation (Falconer and Mackay 1996, 131). Thus the covariance between the phenotypic values (e.g., height) of pairs equals to additive genetic covariance, dominant and epistasis genetic covariance, plus the environmental covariance. A general equation for traditional quantitative methods can be written as follows:

$$Cov(P_1, P_2) = Cov(A_1 + D_1 + I_1 + E_1, A_2 + D_2 + I_2 + E_2) =$$
$$Cov(A_1, A_2) + Cov(D_1, D_2) + Cov(I_1, I_2) + Cov(E_1, E_2) \quad (5)$$

where indexes "1" and "2" represent the two family members for each pair studied. $Cov(P_1, P_2)$ is the covariance between the phenotypic values of one individual with the other.

[1] Or the mean values of their class (e.g., offspring) depending on the particular method used.

A, D, I and E represent additive effects, dominant effects, epistasis effects and environmental effects respectively.

The most commonly used traditional methods for estimating heritability are twin studies. In these studies one already knows that monozygotic twins share almost 100% of their genetic material while dizygotic twins about 50%. The environment is typically divided into the part of the environment that affects both twins in the same way (the shared environment, C) and the part of the environment that affects one twin but not the other (the unique environment, U) (Silventoinen et al. 2003). Hence, in the absence of interaction and correlation between C and U, we have:

$$E = C + U \quad (6)$$

Assuming epistasis effects to be negligible (a common assumption in twin studies), by inserting Equation (6) into Equation (5), we have:

$$Cov(P_{T1}, P_{T2}) = Cov(A_{T1} + D_{T1} + C_{T1} + U_{T1}, A_{T2} + D_{T2} + C_{T2} + U_{T2}) =$$
$$Cov(A_{T1}, A_{T2}) + Cov(D_{T1}, D_{T2}) + Cov(C_{T1}, C_{T2}) + Cov(U_{T1}, U_{T2}) \quad (7)$$

where indexes "T1" and "T2" represent the two twins for each twin pair studied. $Cov(P_{T1}, P_{T2})$ is the covariance between the phenotypic values of one twin with the other.

Because each twin's unique environment by definition is independent of that of the other twin, $Cov(U_{T1}, U_{T2})$ is zero for both monozygotic and dizygotic twins. Given that variance is a special case of covariance when the two variables are identical, and that for monozygotic twins A_{T1}, D_{T1}, and C_{T1} equal to A_{T2}, D_{T2}, and C_{T2} respectively, we can formulate the equation from Equation (7) as follows:

$$Cov_{MT}(P_{T1}, P_{T2}) = V_A + V_D + V_C \quad (8)$$

where $Cov_{MT}(P_{T1}, P_{T2})$ is the covariance between the phenotypic values of monozygotic twin pairs studied.

By contrast, dizygotic twins are expected to share half of their genes, which means that the covariance between the phenotypic values of one twin with the other of dizygotic twin pairs studied $(Cov_{DT}(P_{T1}, P_{T2}))$ is expected to be equal to half of the additive genetic variance, a quarter of dominant variance[2], and all of the shared environmental variance (with $Cov(U_{T1}, U_{T2})$ also to be zero). We have:

$$Cov_{DT}(P_{T1}, P_{T2}) = \frac{1}{2}V_A + \frac{1}{4}V_D + V_C \quad (9)$$

It is classically assumed that V_C in Equation (8) and (9) is the same. That is to say, for both monozygotic and dizygotic twin pairs, it is assumed that the shared environment would act in

[2] For each given gene with two alleles, the possibility that dizygotic twins have the same genotype is one quarter.

the same way if the pair has been reared together.[3] V_C can be cancelled by subtracting Equation (9) from Equation (8). The heritability can then be estimated as follows:

$$h^2_{b_{TS}} = \frac{2\{Cov_{MT}(P_{T1},P_{T2}) - Cov_{DT}(P_{T1},P_{T2})\}}{V_P} = \frac{V_A}{V_P} + \frac{\frac{3}{2}V_D}{V_P} \quad (10)$$

We call $h^2_{b_{TS}}$ broader-sense heritability (the index "b" is for "broader-sense") from *twin studies*, because the resulting estimate (which is about 0.8 for height) provides an accurate estimate of neither H^2 nor h^2, although it is closer to H^2 than to h^2 (Falconer and Mackay 1996, 172). That is to say, it corresponds to a definition of heritability that includes *some* elements of broad-sense heritability but not all of it.

Another often used traditional quantitative method to estimate heritability involves a parent-offspring regression. This method also assumes neither gene-environment interaction nor correlation, the covariance between the height of parents (one or the mean of both) and the mean of their offspring (Falconer and Mackay 1996, 164), equals to additive genetic covariance, dominant covariance (the epistasis covariance is assumed to be small and is not included), plus environmental covariance. Hence, Equation (5) can be formulated as follows:

[3] This assumption might be problematic because monozygotic twins are often treated more similarly by their parents than are dizygotic twins, and monozygotic twins are more likely to share a placenta than dizygotic twins. The difficulty can be mitigated by using adoption twin studies in which the environments for twins are random on average. But large adoption twins' data are exceedingly difficult to get (Griffiths 2005).

$$Cov(P_P, P_O) = Cov(A_P + D_P + I_P + E_P, A_O + D_O + I_O + E_O) =$$

$$Cov(A_P, A_O) + Cov(D_P, D_O) + Cov(E_P, E_O) \quad (11)$$

where indexes "P" and "O" represent the "parents" and the "offspring".

Two assumptions are then made. The first one is that there is no dominant effects transmitted from the parents to the offspring assuming the parents are unrelated (Doolittle 2012, 178), which means $Cov(D_P, D_O)$ is nil. Another assumption is that there is no correlation between the parents' environment and the offspring's environment so that $Cov(E_P, E_O)$ in Equation (11) is also nil. Given that on average, parents share in expectation 50% of genes with their offspring (parents and offspring share half of their genes), it leaves Equation (11) with a result of half of additive genetic variance ($\frac{1}{2}V_A$). Given V_P, h^2 can be estimated straightforwardly.

But the above two assumptions are problematic. First, the assumption of unrelated parents might be violated because of assortative mating in humans resulting in parents to be more genetically similar than two randomly chosen individuals (Guo et al. 2014). Hence, $Cov(D_P, D_O)$ is likely to be non-nil. Second, because the environments experienced by individuals are likely to be more similar within a family line, $Cov(E_P, E_O)$ might not be nil, either. If we take these two factors into consideration, the covariance of the parents and their

offspring is equal to half of additive genetic variance, *plus* a variance term representing effects due to dominance and similarities between environments. This can be written formally as:

$$Cov(P_P, P_O) = Cov(A_P, A_O) + Cov(D_P, D_O) + Cov(E_P, E_O) = \frac{1}{2}V_A + V_{D\&EC} \qquad (12)$$

where $V_{D\&EC}$ represents the variance due to some dominance and environmental correlation effects between the parents and the offspring studied.

The heritability can then be estimated by doubling the parent–offspring covariance in Equation (12) and dividing the total phenotypic variance of the population as follows:

$$h^2_{b_{POR}} = \frac{2Cov(P_P,P_O)}{V_P} = \frac{V_A}{V_P} + \frac{2V_{D\&EC}}{V_P} \qquad (13)$$

For similar reasons as with the heritability estimates from twin studies, we call $h^2_{b_{POR}}$ broader-sense heritability (with the index "b" also being for "broader-sense") from *parent-offspring regression*. Indeed, although it is often assumed that $h^2_{b_{POR}}$ represent h^2 (Falconer and Mackay 1996, 147), the resulting estimate (also about 0.8 for height) is broader than h^2 as it can include a component led by dominance variance and environmental correlation between parent and offspring.

To conclude this section, heritability estimates in both twin studies and parent-offspring regression include an extra term when compared to h^2, but they do not correspond to H^2. For this reason we regroup them under the term h^2_b for "broader-sense heritability", such that:

$$h_b^2 = h^2 + h_{other}^2 \quad (14)$$

where h_{other}^2 is the part of heritability contributed by the extra component(s) representing non-additive variance.

3. **Heritability in GWAS.**

Although any two unrelated individuals share about 99.5% of their DNA sequences, their genomes differ at specific nucleotide locations (Aguiar and Istrail 2013). Given two DNA fragments at the same locus of two individuals, if these fragments differ at a single nucleotide, they represent two variants of a Single Nucleotide Polymorphism (SNP). GWAS focus on SNPs across the whole genome that occur in the population with a probability larger than 1% which are called common SNPs. If one variant of a common SNP, compared to another one, is associated with a significant change on the trait studied, then this SNP is a marker for a DNA region (or a gene) that leads to phenotypic variation. For a polygenic trait like height, if we can detect all the SNPs that associate with it, then all the DNA difference makers that determine height difference can be located.

The development of commercial SNP chips makes it possible to rapidly detect common SNPs of DNA samples from all the participants involved in a study. By using a series of statistical tests, it can be investigated at the population level whether each SNP associates with

that target trait. The choice of the statistical tests depends on the data available as well as the trait studied. For quantitative traits like height, the most common approach is to make an analysis of variance table and assess whether the mean height of a group with one variant at one nucleotide is significantly different from the group with another variant of the same SNP[4] (Bush and Moore 2012). With all the SNPs associated with height being detected, data from the HapMap project, which provides a list of SNPs that are markers for most of the common DNA variants in human populations (Consortium, International HapMap 3 2010), is used to map the associated SNPs with common DNA variants. These mapped DNA variants, to be distinguished from DNA variants that do not affect the target trait, have been called "causal variants" (Visscher et al. 2012).

Based on the readings of SNP chips as well as further independent tests for SNPs, the effects of the associated SNPs (markers for causal DNA variants) on the trait can be calculated. By estimating the phenotypic variance contributed by these SNPs and the total phenotypic variance of the population, the heritability of causal DNA variants can be estimated as the ratio of the phenotypic variance caused by all the associated SNPs compared to the total phenotypic variance of the population (Weedon et al. 2008). Since it is common for biologists to assume

[4] For categorical (often binary disease/control) traits, the association test used involves measuring an odds ratio, namely the ratio of the odds of disease for individuals having a specific variant of a SNP, and the odds of disease for individuals who have another variant at the same locus. If the odds ratio of a common SNP is significantly different from 1, then that SNP is considered to be associated with the disease (Bush and Moore 2012).

that genes are only made up of pieces of DNA, it is thought that the variance obtained from all the causal DNA variants represent exactly the additive genetic variance, and the heritability estimated by GWAS should match narrow-sense heritability (h^2) (Yang et al. 2010; Visscher et al. 2006). However, the assumption that additive genetic effects are solely based on DNA sequences is problematic when faced with the evidence of epigenetic inheritance.

As was mentioned in Section 2, traditional quantitative methods for estimating heritability are based on measuring phenotypic values and genetic relations without reaching the molecular level. The genes are not defined physically, but functionally as heritable difference makers (Falconer and Mackay 1996, 123). In other words, they are theoretical units defined by their effects on the phenotype. With the discovery of DNA structure in 1953, it was thought that the originally theoretical genes were found in the physical DNA molecules. Since then, biologists commonly refer to genes as DNA molecules and this assumption is also made by researchers of GWAS. As Lu and Bourrat claim, this step was taken too hastily. If there is physical material, other than DNA pieces, that can affect the phenotype and be transmitted stably across generations, then it should also be thought to play the role that contributes to additive genetic effects.

Many studies have provided evidence for epigenetic inheritance[5], namely the stable transmission of epigenetic modifications across multiple generations and affect organism's traits

[5] We use the notion of "epigenetic inheritance" in the broad sense that refers to the inheritance

(e.g., Youngson and Whitelaw 2008; Dias and Ressler 2014). A classical example of this is the methylation pattern on the promoter of the agouti gene in mice (Morgan et al. 1999). It shows that mice with the same genotype but different methylation levels display a range of colors of their fur, and the patterns of DNA methylation can be inherited through generations causing heritable phenotypic variations. Epigenetic factors such as self-sustaining loops, chromatin modifications and three-dimensional structures in the cell can also be transmitted over multiple generations (Jablonka et al. 2014). Studies on various species suggest that epigenetic inheritance is likely to be 'ubiquitous' (Jablonka and Raz 2009).

The increasing evidence of epigenetic inheritance seriously challenges the restriction of the concept of the gene in the evolutionary sense to be materialized only in DNA. Relying on traditional quantitative methods, it is impossible to distinguish whether additive genetic variance is DNA based or based on other material(s). Some transmissible epigenetic factors, which are not DNA based, might *de facto* be included into the additive genetic variance used to estimate h^2. This extension of heritable units also echoes to the recent suggestion that genetic (assuming genes to be DNA based) and non-genetic heredity should be unified in an inclusive inheritance theory (Danchin 2013; Day and Bonduriansky 2010).

---

of phenotypic features via causal pathways other than the inheritance of nuclear DNA (Griffiths and Stotz 2013, 112).

To apply the idea that some epigenetic factors can lead to additive genetic effects, the additive variance of them ($V_{A_{epi}}$) should be added to the additive variance of DNA sequences ($V_{A_{DNA}}$) to obtain V_A. Assuming there is no interaction between $V_{A_{epi}}$ and $V_{A_{DNA}}$, we have:

$$V_A = V_{A_{DNA}} + V_{A_{epi}} \quad (15)$$

Inserting Equation (15) to Equation (4) leads to:

$$h^2 = \frac{V_{A_{DNA}}}{V_P} + \frac{V_{A_{epi}}}{V_P} \quad (16)$$

Here we term the first term on the right side of Equation (16) "DNA-based narrow-sense heritability" (h^2_{DNA}), and the second term "epigenetic-based narrow-sense heritability" (h^2_{epi}), we thus have:

$$h^2_{DNA} = h^2 - h^2_{epi} \quad (17)$$

4. **Dissolving the Missing Heritability**.

As we mentioned it in Introduction, since the first successful GWAS was published in 2005 (Klein et al. 2005), there have been a lot of proposals for methodological improvements in GWAS (Manolio et al. 2009; Eichler et al. 2010). Studies have been conducted according to those proposals that permit to obtain higher heritability estimates. Examples include increasing

the sample sizes which has resulted in more accurate estimates (e.g., Wood et al. 2014), considering all common SNPs simultaneously instead of one by one which has increased the heritability estimates of height from 0.05 to 0.45 (see Yang et al. 2010), and conducting meta-analyses which can lead to more accurate results when compared to single analysis (see Bush and Moore 2012). Biologists have also suggested to search for SNPs with lower frequencies than 1% in order to account for a wider range of possible causal variants (Schork et al. 2009).

Aside from these partial improvements, our analysis reveals two reasons explaining away the missing heritability problem: a) In traditional quantitative methods, the heritability estimates include extra terms which are not presented in GWAS; b) In GWAS, heritability is estimated solely from causal DNA variants, while in traditional quantitative methods the additive effects contributed by epigenetic difference (h^2_{epi}) are *de facto* included in the estimates.

These two reasons can be shown formally. Using our terminology, missing heritability (MH) equals to the estimates obtained by traditional quantitative methods (h^2_b) minus the estimates obtained by GWAS (h^2_{DNA}), which are 0.8 and 0.45 respectively in the case of height. Thus we have:

$$MH = h^2_b - h^2_{DNA} \tag{18}$$

Replacing h^2_b and h^2_{DNA} by the right hand side of Equation (14) and (17), we obtain:

$$MH = h^2_b - h^2_{DNA} = h^2 + h^2_{other} - \left(h^2 - h^2_{epi}\right) = h^2_{other} + h^2_{epi} \tag{19}$$

Which means that the missing heritability results from the part of heritability originating from epigenetic factors stably transmitted across generations, plus the part of heritability originating from non-additives factors.

Our point that part of the missing heritability can be dissolved by considering non-additive effects echoes to the claim that almost all GWAS to date have focused on additive effects might be a reason for the missing heritability (McCarthy and Hirschhorn 2008). Although there is not enough data to confirm that non-additive effects do explain away some part of missing heritability, this claim appears again and again in discussions on the missing heritability problem (see for instance Maher 2008; Frazer et al. 2009; Gibson 2010; Kong 2010; Moore 2010). Yang et al. (2010, 565) disagree with this claim and respond that "[n]on-additive genetic effects do not contribute to the narrow-sense heritability, so explanations based on non-additive effects are not relevant to the problem of missing heritability."

We agree with Yang et al. (2010) that non-additive effects do not contribute to h^2. That said, because the heritability estimates obtained from traditional quantitative methods do not strictly correspond to h^2 but include some non-additive elements, non-additive effects cannot be dismissed as irrelevant for the missing heritability problem, though probably they are relevant in a way that both Yang et al. (2010) as well as their opponents did not consider.

5. **Conclusion**.

We have provided two ways in which the missing heritability problem can be explained away. First, heritability estimates from traditional quantitative methods (h_b^2) are overestimated when compared to h^2. The resulting estimates would be smaller if the non-additive elements were eliminated. Second, heritability estimates from GWAS (h_{DNA}^2) are underestimated when compared to h^2 because they do not take into account the additive effects of epigenetic factors behaving like evolutionary genes. The resulting estimates would be larger if epigenetic factors were taken into account. We have voluntarily stayed away from the question of whether heritability should be defined strictly relative to DNA sequences or if it should encompass any factors behaving effectively like an evolutionary gene. Our inclination is that there is no principled reason to exclude non-DNA transmissible factors from heritability measures, but our analysis does not bear on this choice.

References:

Aguiar, Derek, and Sorin Istrail. 2013. "Haplotype Assembly in Polyploid Genomes and Identical by Descent Shared Tracts." *Bioinformatics* 29 (13): i352–i360.

Bush, William S., and Jason H. Moore. 2012. "Genome-Wide Association Studies." *PLoS Computational Biology* 8 (12): e1002822.

Consortium, International HapMap 3. 2010. "Integrating Common and Rare Genetic Variation in Diverse Human Populations." *Nature* 467 (7311): 52–58.

Danchin, Étienne. 2013. "Avatars of Information: Towards an Inclusive Evolutionary Synthesis." *Trends in Ecology & Evolution* 28 (6): 351–358.

Day, Troy, and Russell Bonduriansky. 2011. "A Unified Approach to the Evolutionary Consequences of Genetic and Nongenetic Inheritance." *The American Naturalist* 178 (2): E18–E36.

Dias, Brian G., and Kerry J. Ressler. 2014. "Parental Olfactory Experience Influences Behavior and Neural Structure in Subsequent Generations." *Nature Neuroscience* 17 (1): 89–96.

Doolittle, Donald P. 2012. *Population Genetics: Basic Principles*. Vol. 16. Springer Science & Business Media.

Downes, Stephen M. 2015. "Heritability." In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University.

Eichler, Evan E., Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore, and Joseph H. Nadeau. 2010. "Missing Heritability and Strategies for Finding the Underlying Causes of Complex Disease." *Nature Reviews Genetics* 11 (6): 446–450.

Falconer, Douglas S., and Trudy F. C. Mackay. 1996. *Introduction to Quantitative Genetics*. 4th edition. Longman: Benjamin Cummings.

Feil, Robert, and Mario F. Fraga. 2012. "Epigenetics and the Environment: Emerging Patterns and Implications." *Nature Reviews Genetics* 13 (2): 97–109.

Frazer, Kelly A., Sarah S. Murray, Nicholas J. Schork, and Eric J. Topol. 2009. "Human Genetic Variation and Its Contribution to Complex Traits." *Nature Reviews Genetics* 10 (4): 241–251.

Furrow, Robert E., Freddy B. Christiansen, and Marcus W. Feldman. 2011. "Environment-Sensitive Epigenetics and the Heritability of Complex Diseases." *Genetics* 189 (4): 1377–1387.

Griffiths, Anthony JF., Susan R. Wessler, Richard C. Lewontin, William M. Gelbart, David T. Suzuki, and Jeffrey H. Miller. 2005. *An Introduction to Genetic Analysis*. 8th edition. New York: W. H. Freeman.

Griffiths, Paul, and Karola Stotz. 2013. *Genetics and Philosophy: An Introduction*. Cambridge University Press.

Guo, Guang, Lin Wang, Hexuan Liu, and Thomas Randall. 2014. "Genomic Assortative Mating in Marriages in the United States." *PloS One* 9 (11): e112322.

Jablonka, Eva, Marion J Lamb, and Anna Zeligowski. 2014. *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Revised edition. MIT Press.

Jablonka, Eva, and Gal Raz. 2009. "Transgenerational Epigenetic Inheritance: Prevalence, Mechanisms, and Implications for the Study of Heredity and Evolution." *The Quarterly Review of Biology* 84 (2): 131–176.

Johannes, Frank, Vincent Colot, and Ritsert C. Jansen. 2008. "Epigenome Dynamics: A Quantitative Genetics Perspective." *Nature Reviews Genetics* 9 (11): 883–890.

Klein, Robert J., Caroline Zeiss, Emily Y. Chew, Jen-Yue Tsai, Richard S. Sackler, Chad Haynes, Alice K. Henning, John Paul SanGiovanni, Shrikant M. Mane, and Susan T. Mayne. 2005. "Complement Factor H Polymorphism in Age-Related Macular Degeneration." *Science* 308 (5720): 385–389.

Lu, Qiaoying, and Bourrat Pierrick. Forthcoming. "The Evolutionary Gene and the Extended Evolutionary Synthesis." *British Journal for Philosophy of Science*.

Maher, Brendan. 2008. "Personal genomes: The Case of the Missing Heritability." *Nature News* 456 (7218):18–21.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, and Aravinda Chakravarti. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461 (7265): 747–753.

McCarthy, Mark I., and Joel N. Hirschhorn. 2008. "Genome-Wide Association Studies: Potential next Steps on a Genetic Journey." *Human Molecular Genetics* 17 (R2): R156–165.

Morgan, Hugh D., Heidi GE Sutherland, David IK Martin, and Emma Whitelaw. 1999. "Epigenetic Inheritance at the Agouti Locus in the Mouse." *Nature Genetics* 23 (3): 314–318.

Schork, Nicholas J., Sarah S. Murray, Kelly A. Frazer, and Eric J. Topol. 2009. "Common vs. Rare Allele Hypotheses for Complex Diseases." *Current Opinion in Genetics & Development* 19 (3): 212–219.

Silventoinen, Karri, Sampo Sammalisto, Markus Perola, Dorret I. Boomsma, Belinda K. Cornes, Chayna Davis, Leo Dunkel, Marlies De Lange, Jennifer R. Harris, and Jacob VB Hjelmborg.2003. "Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries." *Twin Research* 6 (05): 399–408.

Turkheimer, Eric. 2011. "Still Missing." *Research in Human Development* 8 (3-4): 227—241.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. "Five Years of GWAS Discovery." *The American Journal of Human Genetics* 90 (1): 7–24.

Visscher, Peter M., Sarah E. Medland, Manuel AR Ferreira, Katherine I. Morley, Gu Zhu, Belinda K. Cornes, Grant W. Montgomery, and Nicholas G. Martin. 2006. "Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings." *PLoS Genet* 2 (3): e41.

Visscher, Peter M., William G. Hill, and Naomi R. Wray. 2008. "Heritability in the Genomics Era—concepts and Misconceptions." *Nature Reviews Genetics* 9 (4): 255–266.

Weedon, Michael N., Hana Lango, Cecilia M. Lindgren, Chris Wallace, David M. Evans, Massimo Mangino, Rachel M. Freathy, John RB Perry, Suzanne Stevens, and Alistair S. Hall. 2008. "Genome-Wide Association Analysis Identifies 20 Loci that Influence Adult Height." *Nature Genetics* 40 (5): 575–583.

Wood, Andrew R, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, and Zoltán Kutalik. 2014. "Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height." *Nature genetics* 46 (11): 1173–1186.

Yang, Jian, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, et al. 2010. "Common SNPs Explain a Large Proportion of the Heritability for Human Height." *Nature Genetics* 42 (7): 565–569.

Youngson, Neil A., and Emma Whitelaw. 2008. "Transgenerational Epigenetic Effects." *Annual Review of Genomics and Human Genetics* 9: 233–257.

Responsiveness and robustness in the David Lewis signalling game

Carl Brusse and Justin Bruner

October 28, 2016

Abstract

We consider modifications to the standard David Lewis signalling game and relax a number of unrealistic implicit assumptions that are often built into the framework. In particular, we explore realistic asymmetries that exist between the sender and receiver roles. We find that endowing receivers with a more realistic set of responses significantly decreases the likelihood of signalling, while allowing for unequal selection pressure often has the opposite effect. We argue that the results of this paper can also help make sense of a well-known evolutionary puzzle regarding the absence of an evolutionary arms race between sender and receiver in conflict of interest signalling games.

1 Signalling games and evolution

Common interest signalling games were introduced by David Lewis (Lewis, 1969) as part of a game theoretic framework which identified communicative conventions as the expected solutions to coordination problems. In recent years, this has informed a growing body of work on the evolution of communication, incorporating signalling games into an evolutionary game theoretic approach to modelling the evolution of communication and cooperation in humans (Skyrms, 2010; Skyrms, 1996).

As the basis for game theoretic modelling of such phenomena, David Lewis signalling games are attractive in their intuitive simplicity and clear outcomes. They are coordination games of common interest between world-observing senders and action-making receivers using costless signals; in contrast to games where interests may differ and where costly signals are typically invoked. In the standard two-player, two-state, two-option David Lewis signalling game (hereafter the '2x2x2 game'), the first agent (signaller) observes that the world is in one of two possible states (state1 or state2) and broadcasts one of two possible signals (signal1 or signal2) which are observed by the second agent (receiver) who performs one of two possible actions (act1 or act2). If the acts match the state of the world (i.e. act1 if state1 or act2 if state2) then the players receive a greater payoff than otherwise.

Most importantly, though, the game theoretic results are unequivocal. There exist two Nash equilibria that are, in Lewis's words, signalling systems where senders condition otherwise arbitrary signalling behaviour on the state of the world, and receivers act on those signals to secure the mutual payoff. The two

systems only differ on which signal gets to be associated with each state of the world[1]. Huttegger (2007) and Pawlowitsch (2008) have shown that under certain conditions a signalling systems is guaranteed to emerge under the replicator dynamics, a standard model of evolution to be discussed further in section 4.

Of course the degree to which Lewis' approach makes sense is the degree to which we have confidence in the interpretation and application of such a highly idealised model to the more complex target systems. The obvious worry is that by introducing more realistic features into the model one might break or significantly dilute previous findings on the evolution of signalling.

Not surprisingly, then, recent work on Lewis signalling games has investigated the many ways in which such de-idealizations could occur. Some deviations from the standard Lewis signalling game include: more and varied states of the world, the possibility of observational error or signal error, noisy signals, partial deviation in interest between senders and receivers, the reception of more than one signal, and so on. Many such concerns are dealt with favourably in Skyrms (2010), and in work by others. For example Bruner et al. (2014) generalizes beyond the 2x2x2 case and Godfrey-Smith and Martinez (2013) and Godfrey-Smith (2015) mix signalling games of common interest and conflict of interest. One complication of the Lewis signalling game (particularly important for our purposes) is that signalling systems are not guaranteed in the simple 2x2x2 case when the world is biased. In other words, when the probabilities of the world being in state1 or state2 are not equal, a pooling equilibrium in which no communication occurs between sender and receiver is evolutionarily possible.

2 Symmetry breaking

The focus here will be with the idealisation that sender and receiver are equally responsive in strategic settings. Senders and receivers (in the evolutionary treatment of such games) are two populations of highly abstract and constrained agency roles: all that signallers do on observing the state of the world is send a signal, and the receivers must act as though the world is in one or other of the sender-observable states. Of those two roles, it is the restriction on receivers which is the more problematic.

Imagine for example a forager sighting a prey animal at a location inaccessible to her, but close enough to be acquired by an allied conspecific (who cannot observe the animal). In this case, it is easy for the first forager to slip into the signalling role and execute it, whistling or gesturing to her counterpart. To play the receiver role, however, the second forager has to actually re-orient their attention (to some degree) and attempt to engage in appropriate behaviour for the world-state the first has observed (e.g. prey is to the east or to the west, etc.).

The Lewis signalling model by design is constrained such that the receiver's actions are limited to just those acts associated with the sender's observed world-states. It is of course sensible to begin inquiry with as simple of a model as possible and consider a limited range of responses to stimuli. However, our point is that it is more plausible to make these idealizations for signallers than

[1]The other two possible outcomes of the game are 'pooling equilibrium', where the receiver plays act1 or act2 unconditionally.

for receivers. Signals are (by stipulation) cheap and easy to send, yet the actions available to the receiver are less plausibly interpreted as intrinsically cheap and free of opportunity cost.

In addition, the informational states drawn on by sender and receiver are also likely to be very different. Any real-life sender's observation of a world state will likely inform their motivations ('we should catch that animal') to dictate a fairly clear course of action ('try to direct the other agent's behaviour'). But all the receiver gets is a whistle, gesture or other signal which (by stipulation) has no pre-established meaning. The experience of observing a strategically relevant state of the world will typically be richer and more detailed than that of observing a strategically relevant artificial signal. All this leads to two concerns. Firstly, asymmetries in the strategic situations are likely to exist between senders and receivers. Receivers are likely to have locally reasonable options available to them other than those relevant to signaller-observed states of the world, and their responsiveness to the strategic situation is therefore less satisfactorily modelled by the strictly symmetric payoff structures of standard signalling games. Call this the structural responsiveness concern.

Secondly, given the likely differences in informational states, goal-directness, workload and opportunity cost implications of sender and receiver roles, we can expect the mechanisms (cognitive and otherwise) which instantiate them to differ as well, quantitatively and qualitatively. This implies that we should not expect their update-responsiveness in any given game to be equal either. Yet the working evolutionary assumption is that senders and receivers update their strategies in an identical manner, modelled using either learning dynamics or replicator dynamics. Call this the evolutionary responsiveness concern.

3 Hedgehog strategies and update asymmetry

The first of these concerns might sound like an argument for abandoning coordination games and moving toward 'conflict of interest' or 'partial conflict of interest' models. However the issue is more specific than this.

The structural responsiveness concern provides parallel motivation to one of Kim Sterelny's (Sterelny, 2012) concerns about Skyrms (2010) use of the Lewis model. Sterelny asks whether the availability of 'third options' on the part of the receiver might undermine the evolution of signalling even when these third options are less valuable than the payoff for successful coordination. As part of a discussion of animal threat responses, he labels this a 'hedgehog' strategy – taking an action which pays off modestly, regardless of the state of the world. To make this concrete, hedgehogs often roll into a ball in response to predators. This is a stark contrast to the more sophisticated behaviour of vervets, who have specific responses to specific threats. Yet the optimal response a vervet takes to one threat – climb a tree when confronted by a leopard – may lead to total disaster when used in response to another threat, such as an eagle. Hedgehogs avoid such outcomes by 'hedging' unconditionally so as to secure a modest payoff. Translated to signalling games, such a gambit may, in many cases, be more attractive than attempting to respond optimally to a signal[2].

[2]It is worth noting here that the 'hedgehog' strategy in this Lewis signalling game is in many ways analogous to the risk dominant 'hare' response in stag hunt games. Playing hare instead of stag allows the agent to avoid disaster, but only guarantees the individual a

This compliments the structural responsiveness concern: receivers (especially) might have other options of value which will stand in competition to those assumed in the standard signalling game. Something like these hedgehog strategies are plausible departures from the idealisation and should be expected on the part of the receiver given a realistic demandingness of the role. The question is whether (as Sterelny suspects) including hedgehog strategies might undermine the robustness of evolution toward signalling systems.

Our second concern pertaining to evolutionary responsiveness parallels a well-known evolutionary hypothesis: the so-called Red Queen effect. In competitive relationships such as predator-prey or parasite-host, the Red Queen hypothesis states that species will be constantly adapting and evolving in response to one another just to "stay in the same place" (Van Valen, 1973). This should also be the case in competitive signalling situations – such as predator-prey signalling systems or courtship displays among conspecifics. Signallers and receivers come to not just update their strategies, but to do so at faster or slower rates depending on the nature of the strategic encounter they are entwined in[3].

It might seem that in David Lewis signalling games (as with games of common interest in general) the Red Queen effect should have no role to play. However any realistic interpretation of the Lewis signalling game makes it plausible to consider asymmetry in evolutionary responsiveness as likely, if not the norm. First, as argued, the precise cognitive mechanisms and procedures employed by senders and receivers are likely to be different. Different systems will admit to different degrees of plasticity and evolvability – and will have a different set of cross-cutting tasks and utilities that will place their own demands upon them. Quick and easy signalling responses will have different pathways of update and adaptation than the (typically) more complex set of systems which appropriate receiver responses require.

The consideration of multiple use or adaptive reuse also makes the Red Queen hypothesis salient: it is wildly implausible that entirely separate cognitive systems would evolve to deal with competitive signalling situations and coordination-style situations. Cognitive structures which underpin sender or receiver behaviour will likely be subject to evolutionary pressures from competitive as well as cooperative situations, and the responsive nimbleness of sender and receiver strategies is therefore not guaranteed to be the same. We should not assume that the evolution of sender and receiver strategies always proceeds at the same pace.

Finally, there is at least some evidence of a basic asymmetry between sender and receiver roles in the literature on great ape communication. For example, Hobaiter and Byrne (2014) stress the great sophistication and flexibility on the receiver side of Chimpanzee gestural communication, while Seyfarth and Cheney (2003) discuss about how greater inferential sophistication on the receiver side is a feature of many primate communication systems. While these findings do

mediocre payoff. Thus the issues and trade-offs associated with the hedgehog strategy are general concerns not confined to just the Lewis signalling games. Thanks to [name redacted for review] for helping us better see this connection.

[3]An example of two groups adapting and evolving at different rates can be found in Richard Dawkin's discussion of his famous Life-Dinner principle (Dawkins and Krebs, 1979). While we expect both predator and prey to adapt to each other, Dawkins claims the prey species will come to evolve at a faster rate than the predator species due to the different selection pressures exerted on both species. Failing to adapt quickly enough for the predator means going hungry for an extra day, while failing to adapt for the prey means death.

not directly support the structural and evolutionary responsiveness concerns, they show that real-life sender and receiver strategies (in our near biological cousins at least) exhibit important differences, suggesting cognitive asymmetries compatible with those concerns.

In summary then, there is reason to consider two structural modifications to the Lewis signalling game as especially salient to the issue of responsiveness: the addition of 'hedgehog' strategies for receivers, and differing rates of change in sender and receiver strategies.

4 The model

The evolutionary model we use as a basis for our analysis is the pure-strategy 2x2x2 David Lewis signalling game, with the two-population discrete-time replicator dynamics.

Exact components of the model include two states of the world (L and R), a world-observing signaller with two possible signals (V1 and V2), and a signal-observing receiver with two possible actions (AL and AR). If the receiver's action matches the state of the world, then both signaller and receiver get a fixed positive success payoff, otherwise their payoff is zero. Signallers and receivers both have four pure strategies available to them (see table 1).

$S1$	Signal V_1 if L and signal V_2 if R
$S2$	Signal V_2 if L and signal V_1 if R
$S3$	Signal V_1 always
$S4$	Signal V_2 always
$S5$	Act A_L if V_1 and act A_R if V_2
$S6$	Act A_R if V_1 and act A_L if V_2
$S7$	Act A_L always
$S8$	Act A_R always

Table 1: Signaller and receiver strategies in the standard 2x2x2 common interest signalling game.

For the evolutionary model, the proportions of the different strategies within sender and receiver populations are initially randomly generated. The fitness of each strategy at a time period t is determined by the composition of the opposing population and the payoff associated with each strategy pairing. The proportion of each strategy at play in the next time period $t+1$ is determined by the standard discrete-time replicator dynamics. For the sender population this is:

$$X_i(t+1) = X_i(t)\frac{F_i}{F_S}$$

where X_i is the ith sender strategy, F_i is the fitness of that strategy and F_S is the average sender strategy fitness. Likewise, for receivers:

$$Y_j(t+1) = Y_j(t)\frac{F_j}{F_R}$$

where Y_j is the jth sender strategy, F_j is the fitness of that strategy and F_R is the average receiver strategy fitness. This is repeated until the populations settle

into an evolutionarily stable arrangement. The update process is deterministic and no randomising or mutations are allowed.

5 Modifications and results

We introduce two novel modifications to this model. First, we add a 'hedgehog' action A_H for the receiver. Second, we allow the rate of generational change of senders and receivers to vary relative to one other. In addition, the bias of nature is also varied, and we investigate the effects these three departures from the Skyrms/Lewis idealisation have on the evolutionary stability of signalling equilibria.

Turning to our first modification, the receiver now has three possible actions upon observing the signal: A_L, A_R, and A_H. As before a success payoff of 1 is received by both players in the case that the receiver plays A_L while the world is in state L, or the receiver plays A_R while the world is in state R. A payoff of zero is received if A_L or A_R is played otherwise. A payoff of H is received unconditionally if the receiver plays A_H, where the value of H is between 0 and 1. The sender has four familiar pure strategies, whereas the receiver now has five (for simplicity we omit conditional strategies involving A_H).

To adapt the earlier forager story, we can imagine the sender and receiver as an egalitarian hunting party, and the game as a situation where the sender remotely observes the location of a valuable prey animal (left or right) and calls out to the receiver. The receiver is initially unable to observe the prey but can choose to go left or go right (catching the prey if they go in the matching direction), or alternatively to abandon the hunt in order to obtain a less valuable resource they do not need help from the sender to acquire (the hedgehog strategy). Varying the prior probability of the world is equivalent to it being in a situation where it is systematically more likely that the prey is to the left or the right.

In the simple unbiased 2x2x2 signalling game, one of the two signalling equilibria is guaranteed to be reached under the replicator dynamics. In our notation, these equilibria are S1-R1 and S2-R2. Increasing the bias of the world (i.e. making L more probable than R or vice versa) will undermine this, with an increasing proportion of populations instead collapsing to pooling equilibria. This will occur when there are initially few conditional signalling strategies in the sender population. In such situations, receivers do best to simply perform the act that is most appropriate for the more likely state of the world. The incentive for senders to adopt a signalling system then disappears and the community is locked into a pooling equilibrium.

Not surprisingly, we found a similar effect with the hedgehog strategy as values of H, the payoff for A_H, becomes significant. The hedgehog strategy R5 is an additional unilateral response, and is able to draw some initial populations away from the signalling equilibria when H is in excess of 0.5 (i.e., the average payoff for 'guessing'). This result, for an unbiased world, is illustrated in Figure 1[4].

[4]Note that the exact range of this effect, including the point at which the effect becomes significant and the y-intercept, are artefacts of the number of world-states and strategies in the model and therefore not general.

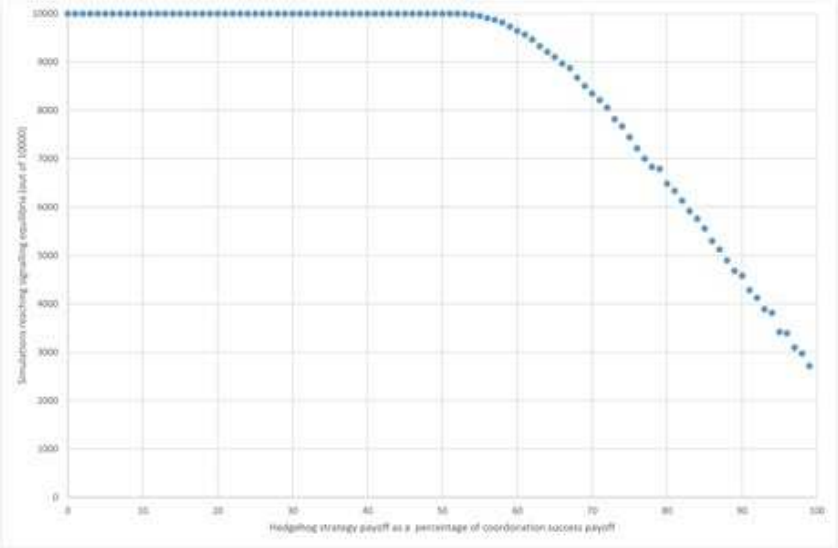


Figure 1: Effect of hedgehog payoff on proportion of signalling equilibria.

We observe a more surprising result when the bias and H are varied in combination. Figure 2 shows the results of varying bias for different values of H. The H = 0 curve has the expected n-shape, with perfect signalling being degraded as world-bias increases away from the mid-point of even bias between L and R. The inclusion of significant (i.e. H ¿ 0.5) hedgehog payoffs decreases signalling at even bias. As nature becomes increasingly biased, however, the proportion of simulations that head to a signalling system does not go down. In fact we observe a 'plateau' followed by a gradual *increase* in the proportion signalling as nature becomes increasingly biased. However, once the bias becomes too extreme, the traditional pooling equilibrium becomes increasingly likely as the payoff associated with simply performing the appropriate act for the more likely state of the world approaches 1. This results in a steep decline in the proportion of simulations that result in signalling systems.

6 Generational asymmetry

We now turn to our second modification of the David Lewis signalling framework in which we introduce a generational asymmetry. We introduced a 'slow-down factor' Z to the replicator dynamics in order control the rate at which sender and receiver populations change over time. Composition of the sender and receiver populations are now governed by the following equations:

$$X_i(t+1) = (1 - Z_S)X_i(t)\frac{F_i}{F_S} + X_i(t)Z_S$$

$$Y_j(t+1) = (1 - Z_R)Y_j(t)\frac{F_j}{F_R} + Y_j(t)Z_R$$

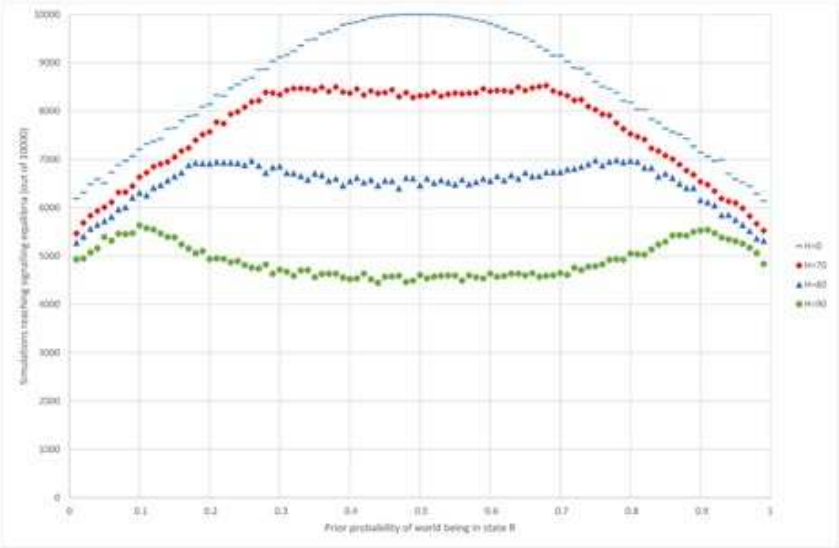


Figure 2: Effect of hedgehog strategy and bias of nature on proportion of signalling equilibria.

Note that when both Z_R and Z_S are zero there is no deviation from the standard replicator dynamics. Rates of changes are slowed as their values increase; for example setting $Z_S = .5$ halves the rate of change for sender strategies. Z_R (alone) being set to 1 means taht the composition of the receiver population would not change over time, and only the sender population would evolve.

The result of introducing this generational asymmetry between senders and receivers is that signalling is more likely when sender strategies evolve faster than receiver strategies. This is illustrated in figure 3, where senders (Z_S) and receivers (Z_R) are slowed down to half and one-tenth speeds (with the other population unaltered) as the bias of nature is varied.

Slowing the evolution of the sender population leads to more pooling because, as before, receivers facing a sender population whose conditional signalling is low will begin to gravitate to the act that matches the more likely state of the world (and the threshold for 'low' is higher at higher bias). This evolutionary trajectory only reverses if conditional signalling increases rapidly enough to tip the fitness balance toward its matching conditional response, before that response is overpowered. Thus signalling becomes quite a remote possibility when bias is high and senders are slow, occurring in less than 10% of simulations for some parameter values. Slowing the evolutionary responsiveness of the receiver population evolves has the opposite effect – as senders will have time to adopt the best separating strategy given the mix of receiver strategies, and the receiver population slowly adjusts and a robust signalling system establishes. By a similar logic, it is easy to see that a quickly evolving sender population also mitigates against the effect of hedgehog strategies.

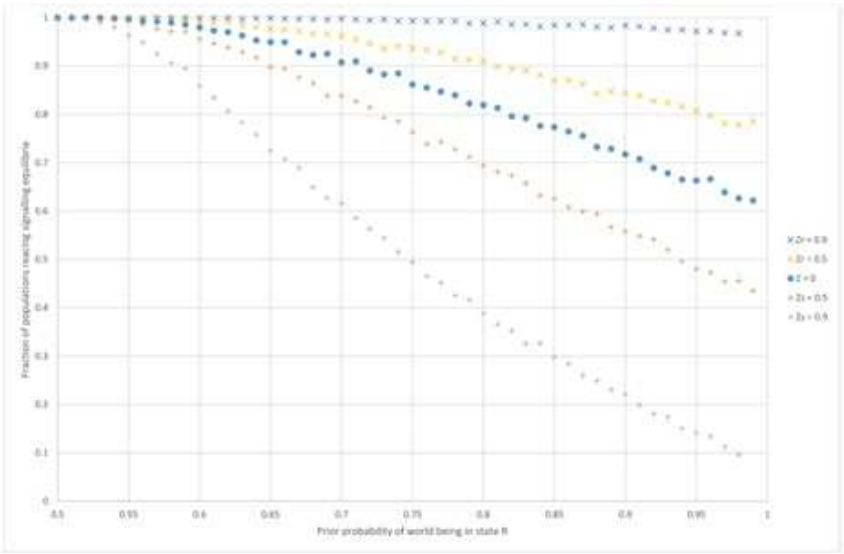


Figure 3: Effect of generational asymmetry and bias of nature on proportion of signalling equilibria.

7 Discussion

We have explored a few well-motivated departures from the highly idealized and simple Lewis signalling game typically considered in the literature. As shown in section 4, breaking the symmetry between senders and receivers often significantly reduces the likelihood that a separating equilibrium emerges. For one, providing receivers with a safe third option which allows them to secure a decent payoff regardless of the state of the world significantly reduces the size of the basin of attraction of the separating equilibrium. Likewise, separating is a remote possibility when receivers outpace senders in the race to adapt.

However the interaction between hedgehog payoffs and bias shows that signalling-undermining effects are not strictly additive. Likewise, the situation is much less bleak when senders evolve at a faster pace than receivers. Interestingly, many scholars in the animal communications literature have noted a similar response asymmetry between sender and receiver in conflict of interest and partial conflict of interest signalling games. For instance, Owren, Rendall, and Ryan (2010) note that senders can easily adapt their signalling behaviour while receivers for the most part have responses to the stimuli produced by senders that are more difficult to change. Thus some have taken to think of signalling as primarily involving the manipulation of receivers by senders.

But this leaves us with an evolutionary puzzle. If there is a conflict of interest between sender and receiver, then what prevents receivers from increasing the speed at which they adapt to the behaviour of the senders? In other words, what explains the absence of an evolutionary arms race between sender and receiver? These are the exact circumstances we would expect the red queen hypothesis to apply. We believe the results of this paper may form the basis of

a novel explanation for this puzzling phenomena. When the interests of sender and receiver are perfectly aligned it is actually in the interest of both parties for the sender population to 'take the lead' and evolve at the faster rate, as doing so ensures the community is more likely to hit upon a mutually beneficial signalling system. When the interests of sender and receiver significantly diverge, however, we would expect this not to be the case since both parties now have reason to adapt at a faster pace than the other.

Yet individuals who routinely interact rarely find themselves playing either common interest or conflict of interest signalling games exclusively. As is well known by any parent, not all signalling interactions between relatives are free of conflict. Likewise, agents whose interests are typically thought to be partially opposed, such as two potential mates, may frequently engage in common interest signalling games in contexts unrelated to mating. The point is that a variety of strategic scenarios can hold between sender and receiver, and there is no principled reason to think all interactions will involve perfect alignment or sizable conflict. If so, then a proportion of signalling interactions between sender and receiver may involve no conflict, a partial conflict, or a full conflict of interest. When the proportion of no or low conflict signalling games is significant, the generational asymmetry result from the previous section may hold to some degree. Both sender and receiver will then profit from the sender population evolving at a faster rate than the receiver population, and receivers do best to limit how responsive they are to senders so as to ensure the emergence of informative signalling systems when their interests do overlap. Thus, while it may appear puzzling as to why a receiver is not more responsive when her interests diverge from that of the sender, this confusion might be resolved when the interaction is put into context.

The robustness analysis considered in this paper has in some sense shown how fragile the evolution of signalling can be. Slightly altering the framework in a sensible fashion leads to significantly different results. While many variants of the baseline Lewis signalling game have been explored by philosophers in recent years, more work is required in order to better assess the prospect of signalling in realistic environments.

8 Acknoweldgements

We thank Kim Sterelny, Ron Planer and the audiences at the Sydney-ANU Philosophy of Biology Workshop and the 2016 Meeting of the Philosophy of Science Association.

9 Bibliography

Bruner, Justin, Cailin O'Connor, Hannah Rubin, and Simon M. Huttegger. 2014. "David Lewis in the Lab: Experimental Results on the Emergence of Meaning." Synthese, September, 1–19. doi:10.1007/s11229-014-0535-x.

Dawkins, R., and J. R. Krebs. 1979. "Arms Races between and within Species." Proceedings of the Royal Society of London B: Biological Sciences 205 (1161): 489–511. doi:10.1098/rspb.1979.0081.

Godfrey-Smith, Peter. 1996. Complexity and the Function of Mind in Nature. Cambridge; New York: Cambridge University Press.

Godfrey-Smith, Peter, and Manolo Martínez. 2013. "Communication and Common Interest." PLoS Comput Biol 9 (11): e1003282. doi:10.1371/journal.pcbi.1003282.

Hobaiter, Catherine, and Richard W. Byrne. 2014. "The Meanings of Chimpanzee Gestures." Current Biology 24 (14): 1596–1600. doi:10.1016/j.cub.2014.05.066.

Huttegger, Simon M. 2007. "Evolution and the Explanation of Meaning*." Philosophy of Science 74 (1): 1–27. Lewis, David K. 1969. Convention: A Philosophical Study,. Cambridge, MA: Harvard University Press.

Martınez, Manolo, and Peter Godfrey-Smith. 2015. "Common Interest and Signaling Games: A Dynamic Analysis." http://petergodfreysmith.com/wp-content/uploads/2013/06/Martinez-GS-paper2-Dynamic-Preprint.pdf.

Owren, Michael J., Drew Rendall, and Michael J. Ryan. 2010. "Redefining Animal Signaling: Influence versus Information in Communication." Biology and Philosophy 25 (5): 755–80. doi:10.1007/s10539-010-9224-4.

Pawlowitsch, Christina. 2008. "Why Evolution Does Not Always Lead to an Optimal Signaling System." Games and Economic Behavior 63 (1): 203–26. doi:10.1016/j.geb.2007.08.009.

Seyfarth, Robert M., and Dorothy L. Cheney. 2003. "Signalers and Receivers in Animal Communication." Annual Review of Psychology 54 (1): 145–73. doi:10.1146/annurev.psych.54.101601.145121.

Skyrms, Brian. 1996. Evolution of the Social Contract. Cambridge University Press. ———. 2010. Signals: Evolution, Learning, and Information. Oxford; New York: Oxford University Press.

Sterelny, Kim. 2012. "A Glass Half-Full: Brian Skyrms's Signals." Economics and Philosophy 28 (01): 73–86. doi:10.1017/S0266267112000120.

Van Valen, Leigh. 1973. "A New Evolutionary Law." Evolutionary Theory 1 (1-30). http://tmtfree.hd.free.fr/albums/files/TMTisFree/Documents/Biology/A

Real Patterns in Biological Explanation

Daniel C. Burnston
Tulane University, Philosophy Department

Forthcoming in *Proceedings of the Philosophy of Science Association*, 2016.

Abstract

In discussion of mechanisms, philosophers often debate about whether quantitative descriptions of generalizations or qualitative descriptions of operations are explanatorily fundamental. I argue that these debates have erred by conflating the explanatory roles of generalizations and *patterns*. Patterns are types of quantitative relationships that hold between quantities in a mechanism, over time and/or across conditions. While these patterns must often be represented in addition to descriptions of operations in order to explain a phenomenon, they are not equivalent to generalizations, because their explanatory role does not depend on any specific facts about their scope or domain of invariance.

Real Patterns in Biological Explanation

1. Introduction

Scientists often claim to have identified patterns in the world. In this paper, I will argue that these patterns are often explanatory in biology, and that their roles in explanation are distinct from the respective roles normally posited for *operations* and *generalizations* in discussion of mechanistic explanation. Operations are types of causal interactions between the parts of a mechanism, described qualitatively. Generalizations are quantitative descriptions of regularities, that normally are taken to involve (at least) two distinct properties in addition to the quantitative relationship. First is *scope*: applicability to a range of cases. Second is *domain of invariance*: insensitivity to manipulations of variables other than those named in the generalization (Woodward, 2010).

Theorists have almost universally equated patterns and regularities, and thus supposed that the explanatory roles of patterns are equivalent to those played by generalizations. For instance, Craver and Kaiser (2013) claim that regularities are "statistical patterns of dependence and independence among magnitudes," (p. 128) and that generalizations describe regularities. Dennett (1991), in his seminal discussion of patterns, calls them a "variety of regularity" (p. 40). Woodward (2010) says that causal relationships are "patterns of dependency" that are "stable or invariant" (p. 291). Most of the literature has followed a similar assumption.

I claim that the explanatory role of patterns is distinct from those of operations and generalizations, and thus that patterns should be considered their own explanatory category.

Patterns, for current purposes, are type-able variations within or between quantities. When biologists cite patterns, they say that a quantity of type X exhibits a particular type of quantitative variation, or that the variations of quantity X stand in a certain type of relation to variations of quantity Y.[1] I will mainly focus on inter-quantity relations here. Often it is important that these relationships occur *across conditions* and/or *over time*—examples include two variations being *proportional* to each other or *in phase* with one another.[1] I will discuss instances of explanation that employ these kinds of relationships, which I have elsewhere (Burnston, 2016) called "explanatory relations." The patterns cited in explaining with these relations are distinct from operations, since they consist in quantitative rather than qualitative types, and since knowledge of the patterns is not fully specified by knowledge of operations. But they are also explanatorily distinct from generalizations, since their explanatory role does not depend on any specific facts about the scope or domain of invariance of the relationships instantiating the pattern.

The initial payoff is simply descriptive adequacy: keeping distinct explanatory categories distinct. I also have a larger target in mind, however. There is currently a considerable amount of debate about whether operations or generalizations are explanatorily *fundamental*—i.e., does one explain the other, or vice versa. "Generalizationists" cite,

[1] While I mean this definition very liberally—the fact that a quantity "increases" in a certain condition is a pattern in this sense—it certainly won't exhaust all colloquial, or perhaps even scientific uses of the concept of a pattern. For instance, one might suggest that one's friend exhibits a negative pattern of behavior without trying to quantify it. Moreover, many patterns are simply statistical facts about a given sample (e.g., noise is "white" only when it has a constant spectral density). Finally, there may also be an infinite number of patterns that are not type-able *by us*. But I'm inclined to think that we need to type a pattern before it can be useful in science, and I will assume that here.

among other considerations, the need for regular quantitative relationships to hold before one can call something an operation (Leuridan, 2010). "Operationists" cite the need for qualitative descriptions of types of relationships in explaining why regularities hold (Andersen, 2011; Machamer, 2004). I think the fundamentality question is, in general, a bad one (cf. Tabery, 2004). In showing that patterns play a distinct role from either operations or generalizations, I hope to suggest that no category is fundamental. This results in a variety of contextualism about explanation.

My strategy is as follows. I will first (section 2) discuss several cases in which biologists explain by representing patterns. In section 3.1, I will argue that this aspect of explanation is distinct from representing operations. I will then (section 3.2) take up a thread in the dialectic between operationists and generalizationists to show that patterns are distinct from generalizations. Some operationists have argued that generalizations are not fundamental for explanation, since we often want to explain in singular or statistically unlikely cases, which involve highly restricted scope and domain of invariance. I will argue that even in cases like these, biologists still need to represent patterns. Hence, operationists are wrong to exclude patterns on the grounds of rarity, and generalizationists are wrong to insist that patterns explain in virtue of having a particular scope or domain of invariance. In both cases, the error is due to equating the explanatory role of patterns and generalizations. I then close (section 4) by suggesting that which explanatory category is most important depends on explanatory context, and thus that there is no fundamentality between explanatory them. As should be clear, my focus is primarily on epistemic concerns. While

the debates about fundamentality discussed above often address both the metaphysics and epistemology of mechanisms, it is productive to keep analysis of these issues separate (Levy, 2013), as I will show below.

I draw my examples from mammalian chronobiology. Chronobiologists study circadian rhythms—roughly 24 hour, endogenously produced physiological rhythms which regulate a large number of processes in the body, ranging from sleep and cognitive abilities, to feeding behaviors, to gene expression. Many organisms have biological "clock" mechanisms within individual cells, which operate on the principle of negative feedback in gene regulation networks. In mammals, the intracellular clock consists in gene regulation between a "negative" loop consisting of the genes *Per* and *Cry* and their respective products (mRNAs and proteins), and a "positive loop" consisting of *Bmal1* and *Clock* and their respective products. In outline, it works as follows. Positive loop proteins bind to E-box promoters on the negative loop genes, activating their transcription. After translation outside of the nucleus, the negative loop proteins dimerize and are translocated back inside of the nucleus, where they bind to the positive loop genes on their own promoters, thus inhibiting their own transcription. As the negative loop proteins degrade, this inhibition is released and the cycle can begin again. With the right rates of transcription, translation, and degradation, these oscillations can occur over a roughly 24 hour period, hence providing a clock signal that can regulate other physiological processes. The clock mechanism is represented in the following diagram.

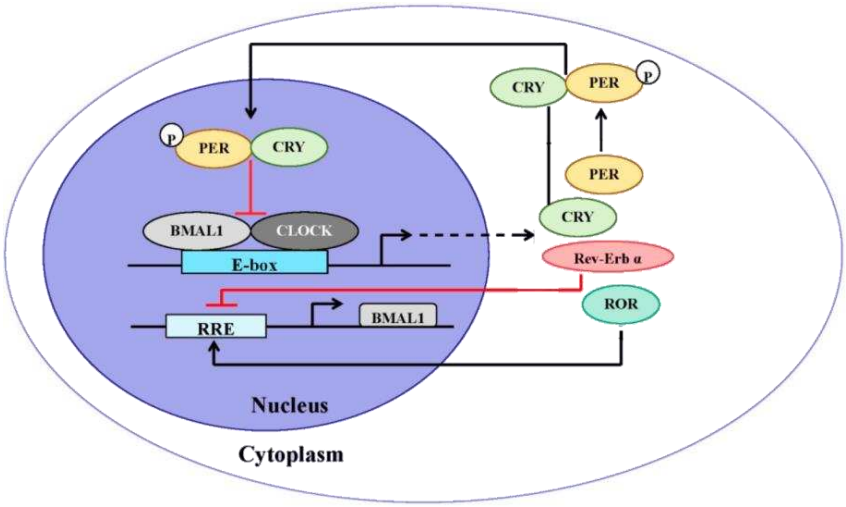


Figure 1. The mammalian intracellular clock mechanism. Modified from Wang, Zhang, Xu, and Tischkau (2014).

In the mechanism, the important parts include the genes and their assorted promoter regions, gene products, the nuclear membrane, etc. The key operations include the activation and inhibition of transcription via selective binding. There are a variety of more complex aspects to the clock mechanism. The products of the positive loop gene *Bmal1* also oscillate, due to a subsidiary feedback loop mediated by *Rev-erb* and *Ror* products. In addition, there are more gene products involved that play support roles, and more types of promoters. Particularly, D-box and RRE (Rev-erb response element) promoters serve as binding sites for a variety of proteins, and each of the promoters can regulate several different genes. Finally, several of the clock genes have *paralogs*—structurally similar genes that serve related functions in the clock.

While the canonical mechanism schema for the mammalian clock, including the parts and operations, has been largely agreed upon since the early 2000s (Zhang & Kay, 2010), investigation into the mechanism has continued—to a significant extent, investigators have turned towards discovering quantitative relationships within the mechanism. In the cases discussed below, I argue that the representation of quantitative patterns over time and across conditions is necessary for explaining certain circadian phenomena. In particular I will focus on temporal patterns regarding *phase* relationships and *proportional* responses in gene networks underlying compensation.

2. Patterns in Explanation

2.1. Phase relationships.

While the mechanistic picture given above is necessary for explaining rhythmicity, it is not sufficient. Several subsequent investigations have shown that it is not only that the mechanism operates according to the schema above that is important, but also that key quantities in the mechanisms bear particular temporal relationships to each other. Looking for these relationships involved measuring and conceptualizing data in certain ways not entailed just by knowing the mechanistic organization.

One such important relationship was discovered by Ueda et al. (2005), who decided to look at the temporal relationships between the activations of gene promoter types *as such*—meaning, regardless of the particular genes that they regulated. Since each type of promoter occurs on multiple distinct genes, analysis of promoters had generally taken a back seat to the study of the genes themselves. However, Ueda et al. showed that the particular

patterns of activity for each promoter type are important for explaining how an entire cell can oscillate in the quantities of its gene products. They first noticed that all of the different activators of a particular promoter type tended to hit their peak expression at similar times, and the same for its repressors. Moreover, for each promoter type—E-boxes, D-boxes, and RREs—there is a distinct phase relationship between their activators and inhibitors. This suggested to the researchers two ideas: (i) that each promoter of a given type is activated *in phase* with other promoters of the same type, even if they regulate different genes; and (ii) that each type of promoter should have a particular phase of peak activation. This is indeed what they found—E-boxes are most active in the morning, D-boxes during the day, and RREs in the evening.

Ueda et al. claimed a functional import for these relationships. Since the clock mechanism consists in a large number of interspersed gene relationships, the phasic regulation of particular promoters across all of the components can keep the many diverse gene interactions on a coherent schedule. For current purposes, however, the explanatory import of the patterns is most clear in a subsequent study by Ukai-Tadenuma et al. (2011). They showed that through very fine-grained manipulation of the *Cry1* D-box, they could manipulate the phase of *Cry1* expression, advancing or delaying it relative to normal D-box mediated expression. Only a phase of D-box-mediated transcription close to wild-type would produce normal cellular rhythms. So, the relative phases of the individual promoter types help to explain how the cell as a whole produces coherent wild-type rhythms.

Importantly, Ukai-Tadenuma did not manipulate the operation performed by the D-box—it still regulated *Cry1* just as normal. Instead, they manipulated the particular temporal pattern of its regulation. So, not only must the particular parts, operations, and causal organization of the mechanism be in place for it to work, but it must *also* have these elements coordinated according to the appropriate temporal patterns. Put simply, if the mechanism did not exhibit this particular set of temporal relations between its promoters, it would not oscillate, and learning this fact was an important addition to the explanation, overtop of the standard mechanism schema given in the clock model. What, then, is the explanatory role being played by the pattern? I suggest that it is adverbial (cf. Burnston, 2016). A mechanistic description shows *what* the operations are and shows the causal organization of their interactions. The representation of patterns shows *how* these interactions are coordinated in their levels and timing to produce quantitative phenomena like rhythmicity. The next example will further illustrate this role.

2.2. Proportionality and compensation.

Baggs et al. sought to study an important phenomenon related to molecular clocks, namely that of *compensation*. In noisy molecular networks, shifts above and below normal quantities of key components are common, but can also be problematic—as shown above, for instance, the clock requires precise temporal coordination of gene product levels in the mechanism. Baggs et al. (2009) showed that compensation in clock mechanisms relies both on their particular mechanistic organization *and* on the particular patterns of change in quantities of gene products as other gene products vary. Their manipulations consisted in

insertion of small interfering RNA (siRNA) into cells in vitro, targeted to specific mRNAs. SiRNA knocks down its targeted mRNAs in a dose dependent fashion, thus allowing for the comparison of responses in varying levels of knockdown. They represented their results in a variety of bar graphs, taken to show the types of responses that were important in implementing compensation. Two are shown below.

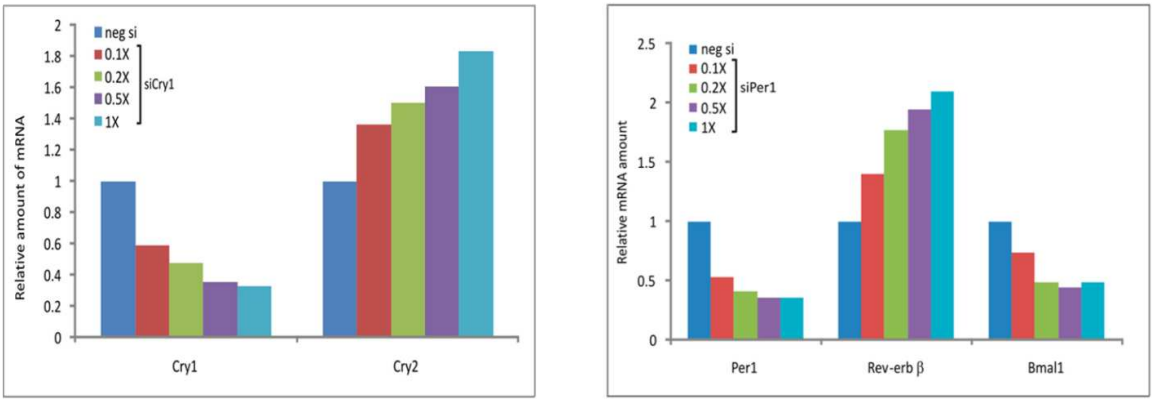


Figure 2. Proportionality patterns in knockdown conditions. From Baggs et al. (2009).

The left panel of figure 2 shows that, with increasing levels of knockdown for *Cry1* mRNA, *Cry2* mRNA *increases*. But not only does it increase, it does so *proportionally*—the greater the knockdown of mCry1, the greater the increase of mCry2. Since *Cry2* is the paralog of *Cry1*, it performs similar operations at similar targets. So, as mCry1 is depleted, the rising mCry2 level results in the overall level of *Cry* influence at its targets remaining the same, thus allowing for the cell's overall pattern of rhythmic gene interactions to continue. Proportional responses are also important in non-paralogous compensation. The right panel shows the effect of mPer1 knockdown on mRev-erbß and mBmal1. *Rev-erbß* is activated by *Per* proteins, and the proteins it codes for inhibit *Bmal1*. When mPer1 levels go down,

mRev-erbß levels go up proportionally. This in turn produces a proportional decrease in mBmal1. The fact that knockdown of mPer1 should cause mRev-erb levels to go up, and that increasing mRev-erb levels should subsequently cause *Bmal1* transcription to decrease, makes sense given the known operations performed by each part: mPer inhibits *Rev-erb*, whose products in turn inhibit *Bmal1*. However, the discovery that each relationship is *proportional* is presented by Baggs et al. as an important further fact in explaining compensation.

It is important for compensation for the following reason: the clock relies on precise interacting levels of inhibition and excitation between the positive and negative loops. Having the levels of one abnormally higher than the levels of the other would wreak havoc on the necessary interplay of inhibition and excitation. As is evident in the right panel, the combined proportional interactions result in a *balance* between the levels of mBmal1 (positive loop) and mPer1 (negative loop), hence keeping the interaction between loops functioning as normal. Knockdowns of other components are compensated for according to similar principles, inducing no loss of rhythmicity elsewhere in the clock.

Proportional relationships, as revealed in the bar graphs, are inherently *patterns of quantitative responses* across knockdown conditions. And, as with the case above, one must represent these patterns in addition to the mechanistic organization to understand how compensation comes about. As Baggs et al. summarize: "the clock network combines these activator and repressor modules with various forms of proportionality to construct relays that generate complex gene expression responses to single gene perturbations" (2009, p. 0570).

So, it is not only the types of causal interactions that occur ("activator and repressor modules"), but also the particular quantitative patterns in which they interact ("forms of proportionality") that explain compensation. This in turn helps explain how functioning rhythms at the cellular level can be maintained despite noisy conditions.

3. Patterns as Their Own Category

3.1. Patterns are distinct from operations.

A category is explanatory when representing it shows, perhaps in part, how the phenomenon of interest comes about. In previous work(Burnston, 2016), I argue in detail that the explanatory role played by representations of patterns is *dissociable* from that played by representations of operations (e.g., in a mechanism diagram). I will only summarize these arguments here, before moving on to discuss the relationship between patterns and generalizations. The key point to note is that in each of the studies above, the parts, operations, and causal organization of the mechanism were already known—neither study extends, revises, or modifies the known mechanistic organization. In each case, however, the researchers discovered and represented a set of relationships between quantities in the system at specific times and/or across specific conditions. As such, knowing the relevant facts about parts and operations *constrains, but does not determine*, all of the relevant facts about the patterns. For instance, in discussing the Baggs et al. case I only focused on linear proportional patterns of responses, but these are not the only possible ones. Baggs et al. also explore several other types, including proportional relationships with fractional coefficients and non-linear responses, which play roles in compensation for other knockdowns. The

point is this: these distinct patterns of relationships are all (epistemically) possible *even given the known operations* performed by each part and the targets they perform them on. So, specifying the parts and operations does not give us all of the information we need to explain. We must also represent quantitative patterns.

3.2. Patterns are distinct from generalizations.

Those who are inspired to consider generalizations as fundamental in explanation often note that mechanisms comprise causal relations, but causal relations of a certain sort, namely ones that are "stable" or "robust" (Leuridan, 2010; Woodward, 2010). A mechanism, the intuition runs, is one that exhibits a stable organization that can produce "regular changes" (Machamer, Darden, & Craver, 2000) in its environment. Hence, mechanisms depend on generalizations instantiated amongst their parts. Those who consider operations fundamental often point to the shortcomings of generalizations for explaining causal relationships *between particulars*. It is the activities of particulars, the intuition goes, that have effects on other particulars, not whether they instantiate some generalization. These relationships can hold even in statistically unlikely or rare cases—in extreme cases, we could want to explain *singular events*, which only happen once. Bogen (2005) and Craver and Kaiser (2013) take this argument to show that explanations do not depend on relationships with a significant domain of invariance or scope, and thus that generalizations only play subsidiary epistemic roles, which help us to access the operations that actually explain.

Patterns of the type I have described, however, are explanatorily distinct from generalizations. The argument involves two claims, one against the generalizationists and

one against the operationists. Against the operationists: representing patterns is necessary for explanation even in cases of minimal scope or domain of invariance. Against the generalizationists: it is the specific pattern in the relationship, not any specific facts about its domain of invariance or scope, which is important for explanation. Presumably, if it were really the case that the explanatory role of a pattern depended on its status as a generalization, then that role would be closely related to how wide a scope the pattern has or how broad its domain of invariance is. The following two simple thought experiments show this not to be the case. The first assesses domain of invariance, and the second assesses scope.

The fragile oscillator. Suppose that we have a system that exhibits the patterns of phase relationships shown in the Ueda et al. study, and thus oscillations amongst the gene products in its molecular clock. But it is highly fragile, meaning that there is an *extremely specific* set of conditions that has to hold in order for it to oscillate. Perhaps the constituent proteins are easily broken apart, or the environment is highly volatile, so that even slight variations in (say) temperature or PH will modify transcription and degradation rates, interrupting the needed patterns and preventing oscillation within the system. One could dress up the example until arriving at a case where the patterns have a *minimum* domain of invariance—that is, in which there is only one set of conditions in which the mechanism will oscillate. In this case wiggling *any* variable other than the ones mentioned in the pattern will prevent the pattern from occurring. If the explanatory role of patterns were based on their having some specific domain of invariance, then they should play a lesser or different

explanatory role in this case than in a case where their domain of invariance is broader. This, I submit, is not the case. When we go to explain how this system works, we will need to mention both its mechanistic organization and the phase relationships between promoters, just as Ueda et al. see fit to do. But if the explanatory role played by representations of the phase relationships is the same in either case, then that role doesn't depend on its domain of invariance.

The lonely compensator. It is important to emphasize here that domain of invariance is distinct from scope. Even if the conditions needed were maximally specific, they could occur in many different instances. To address scope specifically, imagine an opposite case from that above, namely an oscillator that was *so* stable, and existed in *such* an amenable environment, that there were virtually no instances where its gene product quantities varied significantly from their normal (oscillating) values. Now suppose that some cosmically unlikely event occurred, whose only effect was to knock *Per* mRNA quantities away from their normal level. As a matter of historical fact, this has only occurred once, but when it did the system compensated, according to the explanation given by Baggs et al. When giving the explanation for what occurred in this system, if Baggs et al. are right, we will need to posit proportional patterns of the type I described above (along of course, with the standard mechanism schema). Here, ex hypothesi, we have a phenomenon that occurs only once, thus having minimal scope, and yet we still need the representation of patterns in the same explanatory role as in our world where compensation is common. So, the explanatory import of patterns does not depend on facts about their scope.

Both generalizationists and operationists have erred in conflating patterns and generalizations. Against the generalizationists, the explanatory role of patterns does not depend on their having scope or domain of invariance. Against the operationists, they must be represented even in highly specific or unlikely cases. There are likely to be objections from each side. First, generalizationists might insist that, in the thought experiments I've discussed, the patterns *do* have a domain of invariance and a scope; it's just that these are at the theoretical minimum. Hence, they are still generalizations. Operationists, for their part, are likely to suggest that these patterns only "specify key quantities" (to use Bogen's phrase) and that since they do not themselves describe the causal relationships at work, they rely on more fundamental descriptions of operations.

The response to each of these objections is the same: they may make sense as metaphysical claims, but don't tell against the epistemic thesis I am advocating here. I have argued for a particular *explanatory* role for patterns. The cases above show that this explanatory role of patterns remains *the same* regardless of any specific facts about scope or domain of invariance. If a generalizationist wishes to insist that any pattern *must* be a regularity on metaphysical grounds, and is willing to bite the bullet of calling the relationships discussed in the thought experiments regularities, this does nothing to undermine an explanatory distinction between patterns and generalizations. As for the operationist's response, the discussion in section 2 showed that knowing the relevant facts about parts and operations simply doesn't exhaust the explanation. There is a particular role to be played in representing patterns, and this role must be pursued in addition to listing the

parts, operations, and organization. If the explanatory roles are distinct and both necessary, then there is no in principle *epistemic* priority between them (Burnston, 2016). If operationists wish to pursue the fundamentality claim as a metaphysical one, I have no quarrel with them, so long as distinct explanatory roles are kept distinct.

Finally, generalizationists are likely to note that I have leaned on counterfactual reasoning in discussing the role of patterns—i.e., if the patterns *weren't* instantiated, then the phenomenon would not come about. While generalizations are often thought of as grounding counterfactuals, this is different from saying that the explanatory role of a pattern *depends* on its status as a generalization. As the above has shown, we could make the same counterfactual claim regardless of any facts about scope or domain of invariance. For instance, the very same counterfactual holds for proportional relationships in the lonely compensator case as holds in the real world where the scope of proportional relationships is much greater. Again, so long as we are talking about the epistemology of explanation, the role of patterns should be kept distinct.

4. Conclusion: Contextualism and Explanation

I think that the right lesson to draw from the foregoing is that we should distinguish between (i) describing the mechanistic organization of a system, (ii) explaining how a phenomenon comes about, and (iii) generalizing either (i) or (ii). In science, each of these projects is pursued and they are often pursued in tandem; hence they are often run together.[2]

[2] Craver and Kaiser (2013) clearly distinguish between (i) and (iii), but not between (i) and (ii); this is because they miss the distinction between patterns and generalizations, and the important explanatory role played by the

Keeping them distinct, however, allows us to overcome the question of fundamentality by describing the relative roles of operations, patterns, and generalizations in explanation. Aspect (i), obviously, involves discovery and representation of parts and operations. Aspect (ii) often involves aspect (i) *plus* the representation of key quantitative patterns. The thought experiments above show that while aspects (i) and (ii) *can* be extended to ask questions about generalization, they needn't be.

When we *do* turn to generalization, we do so with specific goals and questions in mind. For instance, how widespread phylogenetically is the set of parts, operations, and patterns that implements oscillation? Are other organizations and patterns exhibited elsewhere? At least in terms of mechanistic organization, interacting positive and negative feedback loops between genes is extremely common (although the particular components differ) across a wide range of phyla. This fact about scope is an extremely interesting generalization, since it clues us in to the central importance of circadian timekeeping for all organisms. Equally important, however, is learning the *limits* of these generalizations. One of the major discoveries in chronobiology in the last 15 years is that molecular clocks in cyanobacteria operate on a post-translational mechanism, rather than on interlocking feedback loops of gene regulation (Masato et al., 2005), and hence that the scope of the dual-loop model is limited. Similarly, we could want to know about domain of invariance. For instance, what are the conditions for having a well-functioning clock, and how are they

---

former overtop of describing the relevant parts and operations. Some of what I say about generalization in this section is compatible with Craver and Kaiser's discussion of the distinction between (i) and (iii).

compromised in shift-work disorder, familial advanced sleep phase syndrome, jet lag, and other circadian interruptions? One hypothesis is that jet lag is due to disrupted phase relationships between cellular clocks in two parts of the mammalian suprachiasmatic nucleus (Davidson et al., 2009); hence, in odd lighting conditions the normal phase patterns break down and cannot instantiate wild type behavioral rhythms. These are inherently questions that rely on the generalizations surrounding circadian mechanisms, but the importance of these questions doesn't support the fundamentality of any particular category in giving explanations.

What I want to suggest is that there are simply distinct explanatory contexts, and which category comes to the forefront depends upon the kinds of questions we are asking. For instance, if we are asking what *type* of causal relationship we are analyzing—what parts interact, whether they do so directly, what the results of those interactions are, , etc.—this this predisposes the explanation to invoke operations. When we are interested in how phenomena arise from the operations of a mechanism, attention turns to the interplay of quantities in the mechanism, and thus to patterns and explanatory relations. If we are interested in the robustness of relationships, then scope and domain of invariance, and hence generalizations, come to the fore. This is a distant cousin of contextualisms about explanation that have been advanced before (Van Fraassen, 1983), and while it is not currently a popular way of thinking, I suggest that contextualism is the best way to make sense of the relationship between distinct categories and their relative explanatory roles.

REFERENCES

Andersen, Holly K. "Mechanisms, Laws, and Regularities." *Philosophy of Science* 78, no. 2 (2011): 325-31.

Baggs, Julie E., Tom S. Price, Luciano DiTacchio, Satchidananda Panda, Garret a Fitzgerald, and John B. Hogenesch. "Network Features of the Mammalian Circadian Clock." *PLoS biology* 7, no. 3 (2009): e52-e52.

Bogen, James. "Regularities and Causality; Generalizations and Causal Explanations." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36, no. 2 (2005): 397-420.

Burnston, Daniel C. "Data Graphs and Mechanistic Explanation." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 57 (2016): 1-12.

Craver, Carl F., and Marie I Kaiser. "Mechanisms and Laws: Clarifying the Debate." In *Mechanism and Causality in Biology and Economics*, 125-45: Springer, 2013.

Davidson, A. J. "Visualizing Jet Lag in the Mouse Suprachiasmatic Nucleus and Peripheral Circadian Timing System." *European Journal of ...* (2009).

Davidson, Alec J, Oscar Castanon-Cervantes, Tanya L Leise, Penny C Molyneux, and Mary E Harrington. "Visualizing Jet Lag in the Mouse Suprachiasmatic Nucleus and Peripheral Circadian Timing System." *European Journal of Neuroscience* 29, no. 1 (2009): 171-80.

Leuridan, Bert. "Can Mechanisms Really Replace Laws of Nature?". *Philosophy of Science* 77, no. 3 (2010): 317-40.

Levy, Arnon. "Three Kinds of New Mechanism." *Biology & Philosophy* 28, no. 1 (2012): 99-114.

Machamer, Peter. "Activities and Causation: The Metaphysics and Epistemology of Mechanisms." *International Studies in the Philosophy of Science* 18, no. 1 (2004): 27-39.

Machamer, Peter, Lindley Darden, and Carl F. Craver. "Thinking About Mechanisms." *Philosophy of Science* 67, no. 1 (2000): 1-25.

Nakajima, Masato, Keiko Imai, Hiroshi Ito, Taeko Nishiwaki, Yoriko Murayama, Hideo Iwasaki, Tokitaka Oyama, and Takao Kondo. "Reconstitution of Circadian Oscillation of Cyanobacterial Kaic Phosphorylation in Vitro." *Science* 308, no. 5720 (2005): 414-15.

Tabery, James G. "Synthesizing Activities and Interactions in the Concept of a Mechanism*." *Philosophy of Science* 71, no. 1 (2004): 1-15.

Ueda, H. R., S. Hayashi, W. Chen, M. Sano, M. Machida, Y. Shigeyoshi, M. Iino, and S. Hashimoto. "System-Level Identification of Transcriptional Circuits Underlying Mammalian Circadian Clocks." *Nat Genet* 37, no. 2 (Feb 2005): 187-92.

Ukai-Tadenuma, M., R. G. Yamada, H. Xu, J. A. Ripperger, A. C. Liu, and H. R. Ueda. "Delay in Feedback Repression by Cryptochrome 1 Is Required for Circadian Clock Function." *Cell* 144, no. 2 (Jan 21 2011): 268-81.

Van Fraassen, Bas C. *The Scientific Image*. Oxford: Clarendon, 1980.

Wang, Chun, Zhi-Ming Zhang, Can-Xin Xu, and Shelley A Tischkau. "Interplay between Dioxin-Mediated Signaling and Circadian Clock: A Possible Determinant in Metabolic Homeostasis." *International journal of molecular sciences* 15, no. 7 (2014): 11700-12.

Woodward, James. "Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation." *Biology & Philosophy* 25, no. 3 (2010): 287-318.

Zhang, Eric E, and Steve A Kay. "Clocks Not Winding Down: Unravelling Circadian Networks." *Nature Reviews Molecular Cell Biology* 11, no. 11 (2010): 764-76.

Diagnostics and the 'deconstruction' of models.

Grant Fisher
Graduate School of Science and Technology Policy
Korea Advanced Institute of Science and Technology
fisher@kaist.ac.kr

Abstract: This paper argues that a significant focus in computational organic chemistry, alongside the construction and deployment of models, is the "deconstruction" of computational models. This practice has arisen in response to difficulties and controversies resulting from the use of plural methods and computational models to study organic reaction mechanisms. *Diagnostic controllability* is the capacity of cognitive agents to gain epistemic access to grey-boxed computational models, to identify and explain the impact of specific idealizations on results, and to demonstrate the applicability of computational methods to target systems.

1. Introduction.

In quantum chemistry, providing solutions to the non-relativistic Schrödinger equation for molecules in the ground state from so-called first principles of quantum mechanics has been an acute problem well documented by historians of chemistry (for example Gavroglu & Simões 2012; Park 2009). Recently, attention has been brought to the "computational turn" in quantum chemistry and how it is not merely a matter of technological augmentation but a new discipline of computational quantum chemistry with particular emphasis placed on developments in computational modelling (Lenhard 2014). My aim in this paper is to explore how computational organic chemists attempt to "deconstruct" computational models by diagnosing sources of errors resulting from the use of tractable computational models to study the important organic reactions such as the Diels-Alder reaction. This amounts to modelers attempting to gain epistemic access

to grey-boxed computational processes performed on digital computers by investigating modular computational procedures. These modular procedures include various approximations, tools, and base-line theoretical models. Much contemporary research in contemporary computational organic chemistry is focused on modular computational procedures and their comparative performance with respect to classes of target systems in line with various computational goals as well as constraints like computational speed and cost. While accuracy is a desideratum in computational chemistry a considerable focus of research is the *diagnostic controllability* of modular procedures. Controllability is focused on the applicability of modular procedures to specific classes of target systems and focuses on stability or consistency of result. It is cognitively prior to determining the accuracy of results. Although diagnostics of modular computational procedures is a distinctive practice in contemporary computational chemistry, there are important connections to some of the recent literature on simulations in philosophy of science. Some chemists regard computational modeling as a kind of chemical "experimentation". But even before practitioners can generate simulated "data", what stands out in contemporary computational chemistry is the degree to which practitioners focus on legitimating the application of computational models. I will briefly consider the significance of this issue for computational chemistry and how it might relate to the "verification" and "validation" of computational models.

2. The configuration problem.

In order to compute the activation energies of molecules in organic reactions one has to pay particular attention to the correlation energies associated with different configurations of electrons occupying molecular orbitals. Different interactions between configurations of

electrons can lead to changes in excitation levels crucial when investigating molecules in the activated state. To study these systems, theoretical chemists make use of computational models using digital computers where approximation procedures and idealizations convert equations lacking an analytic solution into a tractable form, resulting in computable algorithms whose outputs permit practitioners to draw inferences about the mechanisms of reactions that are otherwise difficult to access experimentally.

A significant area of early computational organic chemistry addressed pericyclic reactions like the Diels-Alder reaction and the Cope rearrangement. These are reactions are thought to pass through a transition state formed of a closed circle of bonds and are “allowed” when the symmetry of the molecular orbital wave functions corresponding to bonds broken and formed during the reaction is conserved. Michael Dewar was one of the first chemists to propose that digital computers should be used to semiempirically calculate the activation energies and geometries of transition structures for pericyclic reactions. He argued that ab initio methods – that is, calculations supposedly performed from first principles of quantum mechanics – were simply inapplicable to systems of chemical interest because they effectively ignored electron correlations (Dewar & Jie 1992, p. 538). The first semiempirical computational models of pericyclic reactions used approximation methods taking some electron correlations into account while taking many of the core and electron repulsion integrals to be zero so they are not calculated from an explicit Hamiltonian or basis functions. One approximation is called neglect of diatomic differential overlap (NDDO), a powerful semiempirical tool that ignores only the overlap integrals associated with atomic orbitals on different atoms. Other complex integral calculations are replaced by parameters adjusted with reference to experimental data.

Early ab initio calculations in quantum chemistry were based on the Hartree-Fock-Self-Consistent-Field-approach. Electrons are assumed to move in an average potential field comprising the other electrons. One chooses an electron, computes it potential energy, the result is used to compute the next electron, and so on until the calculated potential fields are "self-consistent". Although semiempirical approaches adopted the same base-line model, ab initio methods were restricted to very simple systems without parametrization. But as it became possible to perform ab initio calculations of activation energies using digital computers in the mid-1970s, ab initio and semiempirical computational approaches produced conflicting results. Ab initio models generated results supporting the idea of symmetric transition states in which bonds break and form in unison (synchronously) in a closed circle of bonds. But semiempirical chemists defended alternative reaction profiles tending to suggest asymmetric transition states in which bonds break and form asynchronously.[1] This came to be known as a "dichotomy of methods" because ab initio and semiempirical methods predicted incompatible mechanisms.

Although contemporary chemists tend to play down the seriousness of this dispute, there is recognition that it has done much to shape the character of contemporary computational organic chemistry. That plural approximation approaches to computational modelling resulted in conflicting reaction mechanisms has raised questions about the applicability of various approximation procedures to chemical systems and their reliability given potential errors. Although the computational chemistry community tends to accept the veracity of the ab initio results for pericyclic reactions, that there has been divergence of results for computational

[1] Dewar proposed two alternative mechanisms: reactions either occurred in two distinct kinetic steps via a stable intermediate, or even if in a single step, bonds break and form asynchronously. Both conflict with ab initio results.

models that share the same background theoretical models within molecular orbital theory, but differ in specific model assumptions and the use of parameterization, has resulted in practices of error determination and justification of methods via the "deconstruction" of models.

3. Modular procedures, pluralism, and diagnostics.

Computational modelers aim to generate computational models resulting in algorithms that are tractable in the sense that they are capable of rendering equations that lack analytic solution into computable form. As computational models eschew the intervention of epistemic agents in the computational process, computational models are, as Paul Humphreys (2004; 2009) has argued, at least partially epistemically opaque. In what Humphreys calls "hybrid scenarios", where cognitive agents must "balance the needs of the computational tools with human consumers" (2009, p. 617), not only is this balance to be achieved in terms of computational tractability constrained by computational speed and cost. Consumers aim to deconstruct grey-boxed computational models by cognitively accessing them in order to identify errors resulting from approximations procedures and to determine how they might contribute to the generation of incompatible results. This does not mean that one attempts "full" epistemic access and all errors are eradicated. The idea is that errors resulting from approximations should be "controllable" in the sense to be discussed shortly.

Computational organic chemistry can facilitate epistemic access to computational models due to its methodological characteristics, which include *modularity* and *pluralism*. It is modular in the sense that there are various components parts with specific functional roles used in computational modelling as well as in classifying and organising the tools of the trade. These

modules will include what Humphreys (2004) calls “computational templates”. But modular procedures and tools, unlike templates, will also include techniques that are more specific to computational chemistry. The main kinds of modular tools include minimal or extended basis sets (atom centred functions describing atomic orbitals used to construct molecular orbitals from linear combinations of atomic orbitals) of varying kinds (Slater or Gaussian-type), semiempirical procedures that leave out or approximate the two electron integrals associated with exchange interaction and the correlation interactions using the neglect of differential overlap approximation and correcting the resulting errors using parameters drawn from experiment. These two-electron integrals are a central focus for ab initio procedures and there are various ab initio modular procedures available depending on how practitioners want to tackle electron correlations. Modules include: configuration interaction, Möller-Plesset perturbation theory, multiconfiguration self-consistent field theory, and coupled cluster theory. Some modular tools are employed by both semiempirical and ab initio approximation procedures. For example, a standard theoretical model would be the Hartree-Fock model, employing a mean-field approximation that averages out the effect of electron-electron repulsions. This is a base-line model from which corrections to errors resulting from the mean field approximation are then made iteratively and by augmentation using ab initio or semiempirical procedures. There are also “model chemistries” such as the Gaussian-*n* theories used to benchmark computational results. And as Lenhard (2014) has pointed out, since the 1990s, practitioners have increasingly used density functional theory (DFT) to approximate total energy in terms of the total electron density rather than the wavefunction. The use of DFT is central to the development of computational quantum chemistry.

Computational chemistry is pluralistic because the choice of modular procedure is in part contextual, depending on the extent to which practitioners seek to tradeoff accuracy of computational results for computational speed and cost. Computational studies of the Diels-Alder reaction or the Cope rearrangement are covered in some depth in contemporary research and review articles as well as textbooks where the student and researcher can examine the strengths and weaknesses of what have become "off-the-shelf" computational procedures (see for example Bachrach 2014). This complex modularity has resulted in organization of procedures into hierarchies ("levels of theory") where increasing accuracy demands increasing consumption of resources. The drive towards modularization is important because it represents not only many key developments in computational chemistry, it also facilitates access to partially epistemic opaque computational models. On one level, this is a matter of practitioners diagnosing the sources of errors understood in terms of the contribution made by specific modular procedures to computational results. But this is only made possible by the pooling of comparative studies of computational models and their relative performance across research groups. The general idea of deconstruction is that it is only by first constructing a computational model that delivers solutions comparable to data sets obtained by model chemistries and experimentally determined values that modellers can then go back to the model and diagnose sources of error. The evidence of error is the extent to which results depart from model data or when choice of modular procedures deliver results considered incompatible with computational models making alternative modular choices. Deconstruction is diagnostics. It consists of the collective strategies used to determine the effects of using specific modular components for the study of molecular systems using computational models.

Diagnostics aims at enabling practitioners to isolate, articulate, and quantify error so that they can determine the extent to which computational modules are *controllable*. Here I borrow from Ronald Laymon's (1983, 1987) and Jeffery Ramsey (1992). Both adopt a view of approximation that goes beyond merely assessing approximation validity in terms of the extent to which results depart from experimental data. But there are significant difference between the two authors because for Ramsey "[a]n approximation is an act and not a relation" and will amount to "any methodological strategy which is used to generate or interpolate a result due to underresolved data or deficits of analytic or calculational power" (Ramsey 1992, p. 157). Controllability for Laymon, as I understand it, is essentially being able to give an account of the effect of idealizations of data (Laymon calls these "counterfactual initial conditions") have on the accuracy of predictions such that we are able to seek procedures to "improve" upon them and so improve our predictions. This is central to Laymon's account of confirmation. If one can relax the counterfactuality of the initial conditions used to derive testable consequences from our theories, and the predictions become more accurate, then that theory is better confirmed because it is "monotonic towards truth". A theory is disconfirmed if it does not lead to better approximations (Laymon 1987, p. 211).

My aim here is not engage in the details of the differences between Laymon and Ramsey's respective accounts of approximation and idealization. Both accounts are instructive in that they depart from the idea that approximations should be judged merely in terms of how they might depart from experimental values. But "controllability" in a *diagnostic* sense departs from Laymon's account of confirmation in both its content and its aims (Fisher 2016). First, controllability does not take place in the context of theory-testing but it is nonetheless concerned

with other important epistemic goals in computational modelling. To call a modularized computational procedure *diagnostically controllable* is to claim that one knows how and by how much the distortions introduced will affect computational results given that there are no actual target systems in which the distortions are realized. We want to know what effect, if any, model distortions and other counterfactual assumptions would have in actual cases (this seems to be necessary even before we could say that we can "monotonically improve" upon them). It is important to emphasise that the aim of diagnostic controllability is not necessarily to remove errors nor even to generate accurate results (at this stage at least) but instead to demonstrate that, in spite of the errors, which are to some extent inevitable, computational procedures generate predominantly stable results for a given class of target systems in light the contextual goals of the model users.

For example, ab initio calculations of the Diels-Alder reaction turned out to be controllable because chemists could learn what errors to expect when using the Hartree-Fock model and furthermore that the results would be more or less stable under augmentations to the base line approximation by iterative improvements to the models using some of the modular procedures used to take into account electron correlations. Although improving predictive accuracy is ultimately an epistemic goal in much modelling, controllability demonstrated by relative stability of results under modular iterations can trump accuracy. In early computational organic chemistry, it turned out that while ab initio computations were not always the most accurate, especially in the early attempts to compute activation energies of the Diels-Alder reaction and the Cope rearrangement, they were *controllable* because they more consistently generated results suggesting that the reaction mechanism proceeds in a single kinetic step via a cyclical,

symmetric transition state. Furthermore, DFT approaches tend to reproduce this mechanism. Semiempirical methods tended to be less controllable (it at times generated some notable inconsistencies), and so became considered less *applicable* to studies of, for example, the Diels-Alder reaction and other pericyclic reactions like the Cope rearrangement.

4. Diagnostics and simulations.

Modularization of computational tools and procedures in computational organic chemistry is one way in which epistemic access to grey-boxed computational models is facilitated and enhanced because pluralism promotes a culture of comparative assessment central to the determination of the diagnostic controllability of the modular computational procedures. There are distinctive methods, tools and choices in computational chemistry: to what extent the choice of basis sets will impact on results, how best to approximate electron correlations, what choice of theoretical model to make, to what extent parameters drawn from experiment will contribute to the correction of error and whether the methodological choices involved in parameterization are justified. While much of this is distinctive of computational chemistry, there is much that relates to the existing literature in philosophy of science on computational models more broadly and I would like, in closing, to connect to some recent literature on simulations. One motivation for connecting with the literature on simulations is that for some chemists, for example Michael Dewar and co-workers, computers were "chemical instruments" just as important to chemistry as, say, infrared and NMR spectroscopy but used in a new kind of "experimental" chemistry (Bingham et al 1975, p. 1285). Computational chemistry of organic reactions began because transition structures were inaccessible experimentally and so data was scarce. Furthermore, computers were not just used to crunch numbers. The results or "data" produced by these models

were deployed in model-based inferences about reaction mechanisms and hence chemical dynamics. And what counts as "data" is often derived from model chemistries comprising very accurate computations with large basis sets for relatively small molecules which is used to benchmark modular computational tools deployed in exploring the dynamics of larger chemical systems.

Whether or not simulated data possesses epistemic parity with experimental data is not an issue I can explore here. But I would like to briefly explore some other connections to the simulation literature. Eric Winsberg (2010) argues that the "sanctioning" of computational models where data is scarce depends on possessing model-building principles whose projectability across depends not on the truth of these principles (they are often fictions) but rather their reliability. Diagnostics in computational organic chemistry can be thought of as part of this sanctioning process for computational models in chemistry focusing more on the applicability of procedures given the contextual goals of model users. For example, much of the dispute over ab initio and semiempirical computational methods in the study of the Diels-Alder reaction was whether these methods were applicable to their target system (like the Diels-Alder reaction) and determinations of applicability ultimately closely align to demonstrations of the reliability of computational models.

In any case, applicability is cognitively prior to determining the *accuracy* of computational methods. Here it is useful to draw on a distinction Margaret Morrison (2015) adopts from the simulation literature between "verification" and "validation". Verification concerns whether the equations are correctly solved and is a predominantly mathematical issue whereas validation

concerns how well they computational models represent physical systems. In computational chemistry, and much in line with Morrison's claims, until one has a grip on verification one cannot proceed with validation. A computational chemical "verification" process is at least in part achieved by diagnostics aimed at the identification and quantification of errors and ultimately at determining the applicability of modular computational procedures for the study of classes of target systems. Until applicability is demonstrated, presumably validation cannot be achieved.[2] "Verification" would also include procedural justifications of methods: whether computations are executed in line with model-building goals like top-down ab initio computations or by parameterization from experiment, which are then offered as reasons to support the veracity of computational models.

But there is also something distinctive about diagnostics in computational chemistry in the sense that it might go beyond immediate methodological goals associated with verification. Diagnostics can play an epistemological function in relation to more foundational conceptual issues in chemistry. For example, DFT is known to suffer difficulties in studying non-bonded interactions between molecules. But diagnostic studies of DFT computations of excitation levels in these systems suggest that errors are *unsystematic* because accurate results are possible with some molecules but not others (Peach et al 2008)[3]. In this case diagnostics might reveal aleatory uncertainties associated with errors due to indeterminancy in the physical system.[4] In this case, the uncontrollability of approximations need not reflect poorly on the method. But in the case at

[2] Diagnostics can therefore perform an exploratory function in the sense that demonstrations of applicability are what Axel Gelfert calls a "proof of principle demonstration" in the sense that "a certain type of approach or methodology is able to generate potential representations of the phenomena" (Gelfert 2015 p. 85).
[3] The molecular systems tested include dipeptide, various acenes, H_2CO, HCl, N_2, and CO.
[4] On the distinction between aleatory and epistemic uncertainties, see Morrison (2015, p. 256).

hand, the problem concerns charge transfer excitations in a wide range of molecules, which has prompted researchers to probe the viability of existing conceptions of a chemical phenomenon whose properties they are attempting to describe. This suggests that while diagnostics can correct and change computational procedures, it can also be used to challenge existing descriptions and conceptualizations of the target systems practitioners ultimately aim to represent in the validation stage. Investigations of diagnostic controllability can form a context for the criticism of existing theoretical models and background theoretical assumptions and so is sometimes used to probe the conceptual basis upon which the success of representations are characterized in chemistry. In other words, diagnostics can enter into both "verification" and "validation" in computational chemistry.

5. Conclusion.

Computational organic chemistry has arisen to address the problem of electron correlations in the study of organic reactions such as the Diels-Alder reaction. I have argued that computational differences arising from the use of semiempirical and ab initio procedures has promoted practices aimed at the deconstruction of computational models. This deconstruction is characterized by diagnostics, which focuses on identifying errors arising from plural modular computational procedures (approximations, idealizations, and theoretical models) used in computational modeling. Although accuracy is a desideratum in computational chemistry, a considerable focus of research is diagnostic controllability of modular computational procedures: their applicability and reliability to classes of target systems. This practice, as well as the idea that computers are used to perform experiments, suggests connections to the philosophical

literature on simulations where distinctions between verification and validation seem appropriate but that diagnostic might perform a function in both.

Bibliography.

Bachrach, S. (2014) *Computational Organic Chemistry*. Second Edition. Hoboken, New Jersey: John Wiley & Sons.

Bingham, R.C., Dewar, M.S.J., & Lo, D.H. (1975). "Ground states of molecules. XXV. MINDO/3. An improved version of the MINDO semiempirical SCF-MO method". *Journal of the American Chemical Society*, 97 (6), pp. 1285-1293.

Dewar, M.J.S. & Jie, C. (1992). "Mechanisms of pericyclic reactions: The role of quantitative theory in the study of reaction mechanisms". *Accounts of Chemical Research*, 25, pp. 537-543.

Fisher, G. (2016). "Diagnostics in computational organic chemistry". *Foundations of Chemistry*, 18, pp. 241-262.

Gavroglu, K. & Simões, A. (2012). *Neither Physics nor Chemistry – A History of Quantum Chemistry*. Cambridge, Massachusetts: The MIT Press.

Gelfert, A. (2016) *How to Do Science with Models: A Philosophical Primer*. Springer.

Humphreys, P. (2004). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method.* Oxford: Oxford University Press.

---------- (2009). "The Philosophical novelty of computer simulation methods". *Synthese*, 169, pp. 615-626.

Laymon, R. (1983). "Newton's demonstration of universal gravitation and philosophical theories of confirmation". In Earman, J. (Ed.). *Testing Scientific Theories – Minnesota Studies in the Philosophy of Science*, Vol. X (pp. 179-199). Minneapolis: University of Minnesota Press.

--------- (1987) "Using Scott's Domains to explicate the notions of approximate and idealized data". *Philosophy of Science*, 54, pp. 194-221.

Lenhard, J. (2014). "Disciplines, models, and computers: The path to computational quantum chemistry". *Studies in the History and Philosophy of Science*, 48, pp. 89-96.

Morrison, M. (2015) *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford: Oxford University Press.

Park, B.S. (2009). "Between accuracy and manageability: Computational Imperatives in Quantum Chemistry". *Historical Studies in the Natural Sciences*, 39 (1), pp. 32-62.

Peach et al (2008) "Excitation levels in density functional theory: An evaluation and a diagnostic test", *Journal of Chemical Physics* 128.

Ramsey, J. (1992). "Towards an expanded epistemology for approximations". *Philosophy of Science (Proceedings)*, Vol. 1992, 1, pp. 154-164.

Winsberg, E. (2010). *Science in the Age of Computer Simulation*. Chicago: University of Chicago Press.

Euler's Galilean Philosophy of Science

Brian Hepburn
Wichita State University

Nov 5, 2017
Presentation time: 20 mins

Abstract

Here is a phrase never uttered before: "Euler's philosophy of science." Known as an extraordinary mathematician first, a mathematical physicist second, but never really a physicist — not enough empirical cred — no one has considered whether Euler had a philosophy of science. Even his famed "Letters to a Princess" is described as a somewhat naive parroting of Newton. But Euler is no Newtonian. His philosophy of science borrows from Leibniz, a little from Descartes (in spite of, nay, because of, his critiques of both), but is best seen as continuous with the tradition of a Galilean interpretation of the world as consisting of interacting mechanisms, and the practice of letting the requirements of sound mechanical description and problem solving dictate metaphysics.

1 Introduction

Euler's philosophy of science must be reconstructed from various of his writings, which as a result span a large portion of his life. But a consistent picture does arise. It's main components are a metaphysics, an epistemology, and an explanatory approach. My plan is to describe, briefly, all three, while claiming that they display similarities to Galileo's own philosophy of science.

What I have not found is a "smoking gun"; no explicit acknowledgement by Euler of a debt to Galileo or his approach. But I do think the similarities

are striking. (A supporting argument, not given here, but which I mention in passing, is that just as there are striking similarities between Euler and Galileo, there are stark differences between Euler and either of Newton or Leibniz. Thus, if we are to place Euler in any tradition, it's most plausible that it be Galilean rather than one of these other two contenders.) A brief preview of those similarities are the following:

- a commitment to mechanism and mechanical interaction as an explanatory framework
- a natural science founded equally on matter theory (particularly cohesion) and the science of motion
- seeing mathematics as descriptive of mechanisms, rather than merely motions

The last claim I think will sound most controversial, given interpretations of Galileo's law of falling bodies, for instance, as merely descriptive of the motion and not offering a causal explanation at all. I think that's a misinterpretation, that Galileo did seek a causal interpretation of both the time-squared law and the uniform acceleration law. A number of the diagrams on Galileo's working folios seem to be attempts to derive these laws from natural circular motion. Nonetheless I'm willing to concede that, although Galileo did not see math as describing mechanisms or causal relations between quantities (how could the square of time be a cause of a body's position?), Euler does see the relations that way, at least the *appropriate* relations, the *principles* of mechanics. Euler's position would then represent both an innovation and maturation of Galileo's philosophy of science.

In what follows, I will demonstrate Euler's side of the connection: the role of mechanism in his natural philosophy, his views on the essence of material bodies, and commitment to all causes of change in motion through contact.

2 The organization of Natural Science

In addition to the formal domains of mathematics and logic, Euler distinguishes three different areas when speaking about the study of nature. Natural science is his broadest category, the explanation of causes which affect material bodies.

> Natural science is a science that aims to explain the causes of change that occur to material bodies. (E847, p 1)

But within Natural Science, more specifically, are statics and mechanics, for bodies at rest or in motion, respectively, and metaphysics, which is basically an exercise in interpretation for Euler; viz. it is reasoning about the properties the world must have given the truths of statics and mechanics. (For a discussion of statics and mechanics, see the preface to Euler's *Mechanica*, E015; for the characterization of metaphysics:

> Metaphysics ... is occupied with the investigation of the nature and properties of bodies ... knowing [the truths of Mechanics] will serve as a guide in these thorny investigations. (E149, p2)

See also E200, with the title "Essay on a metaphysical demonstration of the general principle of equilibrium".)

The importance of the organization is the hierarchy implicit it in. The main activity of Natural Science is mechanics for Euler, the study of how forces affect the motion of bodies. But mechanics is embedded within a broader program, which recognizes both the aim of mechanics (demonstrating the causes of change) and the conditions necessary for the possibility of that program (the metaphysics of material bodies.)

3 Criterion of understanding

Mechanics is a mathematical discipline, but since the point of mechanics is explanation of the causes of motion, the mathematical language used within mechanics must meet a criterion of understanding. Euler describes that criterion in the Preface to his *Mechanica*.

> But in all writings which are written without analysis it happens most in Mechanics that the reader, although convinced of the truth of those things which are put forward, nevertheless does not achieve clear and distinct knowledge of them, and so can barely solve the same questions by his own devices when they are altered even a little, unless he engages in analysis and explicates the same propositions using an analytical method. This often happened to me when I began to read through Newton's *Principia* and Hermann's *Phoronomia*: although I seemed to myself to have understood the solutions to many problems, still I could not solve other problems that differed even a little.

Only the analytic approach to mechanics provides a proper understanding of mechanics, and that is a proper understanding of the causes of the various ways the motion of a body can be effected. The analytic calculus has

certain representative affordances which enable easy extrapolation from one solution to another, physically similar, situation.

The feature which makes this so is the ability, in analytic calculus, to represent and operate on functions. Those functions represent the mechanical connections between bodies, equilibrium relations among quantities and changes in quantities. This is a representation more general than merely a mathematical description of a body's motion, or a trajectory in time.

An important class of functional relations are those which describe constraints among the parts of different types of bodies. Euler has a plan for mechanics, exemplified by his plan, albeit, not completed, for his Mechanica. It beings by considering points and the effects of forces on their motions; next are rigid bodies, then fluids, and finally gasses. The nature of these bodies means that the effects of forces on those bodies is different. They are modified by the internal mechanical connections of those bodies.

Functions describe the internal mechanical connections. They also express general properties of bodies. A function describes abstract relations among variables which represent *kinds* of physical properties. An actual existing body is one fully determined, i.e. with specific values for all it's variables.

> If the essence is stipulated in its totality, there arises a single body, containing nothing indeterminate. Such a body is representative of all material bodies really existing as part of this world, since nothing can exist in reality that is not fully determined. Whilst the essence of material bodies in general is subject to few stipulations, the particular types of body, and the individual bodies belonging to each type, arise when the constraints placed on the essence are complete, so that nothing is left indeterminate. (E842, S.8)

Constraints arise through the nature of the body (whether solid, elastic, liquid or gas) and through interactions with other bodies.

There is thus, a close connection between Euler's metaphysics and his mechanics, through functions — and through functions to explanation and equilibrium expressed as mechanical connection.

4 The general properties and essence of bodies

In E847, "An Introduction to Natural Science", Euler considers four general properties of all material bodies, finally arriving at impenetrability as the essence of matter.

> Every material body must occupy in space a particular location, and it is impossible for two bodies to be at the same location at the same time. (E842, p20)

A body could not be said to occupy space unless it was impenetrable. If another body could pass through the location of a body, it could not be said to occupy that space.

The other properties Euler considers as candidates for essence all depend on impenetrability because they, too, depend on the ability of a body to occupy space. Those other properties are extension, mobility, and persistence in a state of motion.

A further distinction is made between what Euler calls coarse matter and subtle matter. The latter is the ether, which differs from coarse matter in that, while it is impenetrable, it is also elastic. Ether is the medium in which light travels, and is also responsible for gravity. Tension in the ether causes there to be a pressure on bodies. The pressure weakens the faster the ether moves, in a direction orthogonal to the motion of the ether. And, since the ether moves faster closer to massive bodies, the difference in sideways pressure causes bodies to move towards heavier ones in the required $1/\mathrm{R}^2$ relation.

The pressure of the subtle matter is also responsible for cohesion of matter, as the ether invades the pores of bodies and surrounds bodies. Again, through the elastic pressure of the ether, bodies cohere, the pressure outside of a body being greater than within. There is no void, for Euler. All of space is filled with either coarse or subtle matter.

5 Impenetrability is the cause of all change

Given that the world is a plenum, Euler can ascribe all change to change through contact. Gravitational attraction is not a property of bodies. There is no gravitational mass, there is no gravitational force.

> 140. Gravity arises from the unequal pressure of the aether, which increases with increasing distance from the earth; therefore the bodies are more strongly pushed towards the earth than

> away from it, and the net excess of these pushing forces is the weight of the body.

All change is through contact. When two bodies come in to contact it is impossible for them to persist in their current state.

> 88. A body is pushed or pressed by others when, because of its impenetrability it is in their way, so that they cannot remain in their state; and through this push or pressure the state of the body itself is altered. From these circumstances originate all forces that act on bodies. (E842, §88)

Each body, in attempting to maintain their state, acts to change the state of the other body. Only in this way is persistence, which Euler says is the proper notion of inertia, properly considered a Force. A Force is the cause of a change. No force is required, therefore, to persist in a state of motion. Persistence is a cause a change of motion in other bodies.

Equilibrium is crucial to the explanation of change through contact. A longer explanation is given in the paper where Euler considers the controversy over vis viva. That paper is E082, titled "On the force of percussion and its true measure", published in 1746. In that paper, Euler gives his account of the proper understanding of what others call the force of inertia. His idea is that when an obstacle is encountered so that a body can no longer maintain its present state there will then be inertia in excess which is no longer consumed within the first body but rather acts to change the state of the other body.

> For as long as the Body remains in the same state of movement or of rest the force of inertia is consumed by conserving its state and consequently is deployed entirely to the inside of the body, without producing anything to the outside. But when external obstacles prevent the body from persevering in its state, so that the force of inertia cannot produce its effect to the inside of the body, then it is deployed to the outside and acts on the external obstacles so that the loss that its effect suffers in the body is exactly compensated by its external action. (E082, p.26)

Any difference in mass between the two bodies will mean that the inertia results in a different amount of speed being given up to the other body, in inverse proportion to the masses. Euler's conception of inertia thus constitutes a physical explanation of the conservation laws which govern collision.

If we construe this operation as an equilibrium or balance exchange we can understand this as a mechanical explanation, relying on analogy with a

lever. [1] This allows a broader class of mechanisms, by allowing functional, equilibrium relations between new quantities such as action. But Euler's conception does not include action at a distance. It is a balanced exchange through the pressure of contact and contact only, caused by the impenetrability of bodies and the force of their attempt to persist in their state of motion.

6 Summary

Euler explains all change in Natural Science through the cause of inertia. The metaphysics required to make this explanation possible makes impenetrability the essence of material bodies. Other general properties, require impenetrability. In particular, when bodies contact, because of their impenetrability, they cannot persist in their given state of motion. The inertia of a body therefore manifests as a force which changes the state of the other body. The motion they lose in this way is equal to the change in motion of the contacted body, and vice versa. Equilibrium is maintained through a balanced exchange of motion.

[1] Machamer, McGuire and Kochiras have recently argued for this way of generalizing the mechanical philosophy as a way of making even action at a distance a mechanical operation. This I think goes a little too far, and it's hard to see why it wouldn't simply make every cause and effect explanation a mechanical one (Aristotle becomes a mechanical philosopher).

Trade-offs between Epistemic and Moral Values in Evidence-Based Policy

Donal Khosrowi
Durham University
30 October 2016

Abstract: I examine the role and relationship of epistemic and moral values in the Evidence-Based Policy (EBP) paradigm. I argue that several epistemic values that play a crucial role in shaping standard EBP methodology stand in a trade-off relation with certain kinds of moral and political values. This is because the outputs afforded by standard EBP methods are insufficient for the pursuit of moral and political values that require information about the distribution of individual treatment-effects among agents in a population. I examine a potential reply to this standard concern, and argue that the changes to standard EBP methodology required for rendering research outputs informative about the distributive consequences of policy typically involve the sacrifice of several key EBP epistemic values at once. I expand on the implications of this trade-off for value-freedom and -neutrality in EBP.

Keywords: **Evidence-Based Policy, epistemic values, non-epistemic values, trade-off**

1. Introduction

Evidence-based policy (EBP) is the call that public policy formation should be informed by high-quality empirical evidence for policy effectiveness from randomized controlled trials (RCTs) and meta-analyses. In emphasizing the superior epistemic credentials of these methods, EBP advocates seek promote several epistemic values such as rigor, unbiasedness, precision and the ability to obtain causal conclusions about policy effectiveness.

In what follows I argue that these epistemic values stand in a *trade-off relation* with a wide range of moral values that policy-makers may be interested in pursuing. Specifically, I argue that standard EBP methodology severely complicates policy makers' ability to pursue moral values such as equality or priority for the worst-off. This is because standard EBP methods are not informative about the *distributive consequences* of policy (see e.g. Manski 2000). This is a substantive shortcoming, particularly when we have reasons to suspect that a policy will render some agents worse off. Yet, since the evidence typically afforded by EBP methods is uninformative on such distributive consequences, it is differentially useful for the pursuit of different moral and political values, specifically utilitarian vs. non-utilitarian values. I argue that this challenges both value-freedom and neutrality in EBP.

The contents are organized as follows. In Section 2 I offer a sketch of the epistemic values involved in EBP as well as whether and how EBP involves ideals of value-freedom and neutrality. In Section 3 I expand on the epistemic challenges that standard EBP methodology faces with respect to generating information about the distributive consequences of policy from RCTs. I discuss how this problem can be addressed by performing subgroup analyses and expand on some of the challenges that this method faces. I also comment briefly on the extent to which these issues have been anticipated and addressed in the extant EBP literature. In Section 4 I give my argument for the trade-off relation between basic EBP epistemic values and moral values that are sensitive to the distributive consequences of policy. I expand on how this trade-off challenges both value-freedom and neutrality in the EBP paradigm. Section 5 concludes.

2. Epistemic Values in EBP

Before I sketch the central epistemic values in the EBP paradigm and how they relate to EBP methodology, it is important to note that there is perhaps no univocally accepted set of epistemic values common to all activities under the EBP heading. More fundamentally, it may be contested whether there is something like a unified EBP paradigm at all. The EBP movement, particularly as it changes over time and in response to various criticisms, is difficult to precisely demarcate as a unified paradigm with distinctive and invariant objectives, methods, underlying epistemic value presuppositions and so forth.[1]

Even so, it is not entirely misleading to think that there is a kernel of epistemic values that are common to a broad variety of activities under the EBP heading. It is this kernel of values that I focus on. These values are not coextensive with traditional epistemic values in the context of theory choice or appraisal such as those offered by Kuhn (1977). Instead, for empirical paradigms such as EBP it seems more plausible to consider values that concern the production of treatment-effect estimates. More specifically, the values that I focus on are *rigor*, *unbiasedness*, *precision* and *the ability to obtain causal conclusions* on grounds of EBP evidence. I consider these values to be prima facie uncontroversial instances of purely epistemic values that seem to be shared among many EBP practitioners. While it is

[1] In addition to this caveat, it is important to note that the construal of Evidence-Based Policy I consider here is somewhat constrained in that it prevalently focuses on the so-called *treatment-effects literature* as instantiated in e.g. econometrics and evidence-based economics, evidence-based medicine and educational research. The distinctive characteristic of this literature is its predominant focus on experimental and quasi-experimental methods to estimate treatment effectiveness. This is considerably narrower than a construal of evidence-based policy as policy that is informed by any empirical evidence rather than only specific kinds of such evidence. I thank Erin Nash for raising this important point about the scope of Evidence-Based Policy.

not always clear what these values specifically consist in, my argument is sufficiently broad to cover most plausible construals that they permit.[2]

The values that I focus on are central to EBP in the sense that they jointly give rise to (and are promoted by) standard EBP methodology, i.e. a set of salient methodological principles that seem to be shared among proponents of the paradigm.

For instance, EBP methodology specifically focuses on certain epistemic targets, i.e. causal conclusions about policy effectiveness. Moreover, EBP methodology is premised on principles concerning the relative desirability of certain kinds of evidence, e.g. by emphasis of the superiority of experimental and quasi-experimental contra purely observational evidence. Finally, EBP methodology emphasizes the relative ability of different methods with respect to generating desirable kinds of evidence; again by focusing on RCTs (and quasi-experimental designs) as opposed to observational studies.

Together, these methodological principles mediate between epistemic values and methods in the sense that EBP methodology promotes values such as rigor, unbiasedness and causal inference *in virtue of* recommending the use of RCTs.

2.1 Value Neutrality and Freedom in EBP

Aside from the identification of crucial EBP epistemic values, it is important to consider whether EBP involves some ideal of value-freedom and/or neutrality. Similar to the issue of identifying key EBP epistemic values, it is not obvious that EBP proponents in general pursue any specific ideal with respect to value-freedom and neutrality.

Even so, it seems that the EBP paradigm rests on a relatively broad axiological presupposition that a *division of labor* with regard to settling normative issues of what values policy should promote and settling factual issues of what are effective means to promote these values is possible. In other words, EBP proponents seem to assume that agreement on the desirability of policy outcomes can be *separated* from the production of evidence speaking for the efficacy and effectiveness of policy in realizing these outcomes.

[2] There may be several additional candidate epistemic values that appear to play prominent roles in shaping EBP research but are not considered here. One such candidate is *generality*, where the principled aim is to establish general claims about the causal efficacy of intervention-types that are robust across time, environments, populations, and individuals. This value seems particularly relevant for extrapolation of causal claims to novel targets; an issue that is related to, but epistemologically distinct, from the issue of welfare analysis of extant interventions that I focus on. I thank Heather Douglas for proposing this additional candidate value at the "Science, Values and Democracy" workshop in Tilburg, NL.

This broadly parallels traditional ideals regarding the role of non-epistemic, moral values in economics, where economists have frequently invoked the metaphor of *economists as social engineers*, who provide factual answers to policy questions *independently from* and typically after policy makers have settled issues concerning the relative desirability of social outcomes (cf. Hausman and McPherson 1996). While I am not claiming that EBP proponents subscribe to this particular ideal, EBP methodology seems to presuppose at least that *some such* division of labor is possible. Let me expand on what this suggests for the role of value-freedom and neutrality in EBP.

First, it seems plausible that many EBP proponents pursue some ideal of value-freedom in the sense that non-epistemic values are generally not and should not be involved in shaping the conduct and outcomes of EBP research *internally*. For instance, while non-epistemic values may be involved in selecting outcome variables of interest, or may act as constraints on whether conducting RCTs is morally permissible, non-epistemic values are generally not and should not be involved in the choice and application of methods once these issues are settled. For instance, the choice between RCTs and observational studies, or the interpretation of estimands obtained from such studies, should not vary with respect to researchers' preferred conclusions about the desirability of the policies under scrutiny. These *internal* aspects should be guided by epistemic values alone.

Second, I consider EBP proponents to pursue *some* version of value-neutrality in the sense that the outcomes of EBP research are intended to be value-neutral insofar as they should not, and generally do not issue unconditional normative claims about the relative desirability of social outcomes or the interventions that promote them. At most, *if* there are normative claims issued in the dissemination of EBP research, these claims take the shape of *hypothetical imperatives*, i.e. normative claims that are conditional on some substantive value presupposition but do not endorse this value presupposition as such.

In order for EBP research to maintain value-neutral, the adequacy of such presuppositions speaking for the desirability of some social outcome must be settled *independently* from (and perhaps prior to) generating information about the relative effectiveness of different interventions in producing the outcome. If such independence is achieved, then *even if* EBP research sometimes issues normative claims, these claims are still value-neutral since they remain non-committal on the adequacy of the substantive moral value presuppositions involved. This issue is left to policy-makers to settle.

With this brief exposition in mind, let me focus on the underlying reasons for why the epistemic challenges involved in generating information about the distributive consequences of policy yield a trade-off between the epistemic values outlined above and non-epistemic values such as equality and priority for the worst-off.

3. Treatment Effect Heterogeneity

Public policy interventions almost invariably affect agents in heterogeneous ways. Consider for instance the case of *microfinance programs*, i.e. programs that supply microcredits to agents who lack access to capital markets. Let us grant for the moment that at least some of these programs may be successful in generating positive long-run welfare consequences for target populations, e.g. by increasing average household endowment or private investment. Even so, behavioral response to microfinance access often differs significantly between agents (cf. Banerjee et al. 2015)[3]. Some agents, e.g. those whose otherwise successful entrepreneurial efforts are inhibited by inadequate access to capital markets, may significantly benefit from such programs. Yet, other, economically less sophisticated agents may be driven into debt traps by pursuing unprofitable business plans and taking up high-interest loans in order to repay initial program loans.

Such heterogeneity in individual treatment effects is predominantly attributable to differences in the causal mechanisms involved in the production of the outcomes of interest or the individual-specific realizations of variables that figure in these mechanisms. This means that the mechanisms connecting treatment and outcome variables of interest typically involve various factors other than treatment that affect the causal relations between treatment and outcome in different ways. For instance, the mechanisms that causally relate microfinance access and eventual welfare consequences for target agents are plausibly mediated and moderated by an extensive battery of factors such as entrepreneurial ability, education, prior business ownership, pre-intervention budget constraints, business plan feasibility etc. These and other factors jointly moderate or mediate the causal effect of treatment on outcome, and agents will typically differ with respect to their individual-specific realizations of these factors as well as whether and how these factors are involved in the individual-specific mechanism that govern the production of the outcomes of interest. As a consequence of such differences, individual treatment effects with respect to one and the same intervention will typically differ between individuals.

This kind of causally relevant heterogeneity is likely to obtain in many areas traditionally targeted by EBP, e.g. in educational policy, where students may respond differentially to educational initiatives as a function of initial ability; in economic policy where policy outcomes may differ significantly between industries, individual firms and other agential units; and in public health and development economics, where agents' response to programs such as bednet distribution might exhibit substantial heterogeneity as a function of agents' basic needs or epidemiological knowledge.

[3] Cited with permission from the authors

As these stylized facts indicate, heterogeneity among agents' response to treatment is ubiquitous in several key areas targeted by EBP. Yet, the issue of heterogeneity has only recently attracted attention from EBP proponents (in marked contrast to evidence-based medicine, see e.g. Oxman and Guyatt 1992 for an early treatment). This is surprising because heterogeneity is responsible for one of the most basic inferential challenges that EBP faces, i.e. the problem of extrapolating experimental results from study populations to eventual policy targets. Let me expand on some technical background to explain why this is the case.

3.1 Heterogeneity Information from RCTs

Technically, treatment effect heterogeneity is the systematic variation in the sign and/or magnitude of individual treatment effects among agents subject to a given intervention. In a potential outcomes framework (Rubin 1974, Holland 1986), given an outcome of interest Y, the individual treatment effect (ITE) for individual i is the difference between her potential outcome $Y_i(1)$ given the treatment and her potential outcome $Y_i(0)$ in the absence of treatment, other things being equal. Since only one of the two values of Y_i can ever be observed, ITEs are in principle unobservable magnitudes.

RCTs can be considered to remedy this inferential dead-end at least to some extent by permitting the estimation of *average treatment effects* (ATEs) instead of ITEs. This is achieved by randomization of confounding factors and treatment moderators and mediators[4] through random assignment of subjects to experimental and control conditions and multiple blinding of trial participants, those administering treatment and those recording and interpreting outcomes. Provided that randomization (and blinding) are successful in that the net effects of confounders and moderators (as well as their interactions) are approximately balanced between treatment and control groups, an ideal RCT can help obtain a consistent estimate of the ATE by taking the difference in means of Y for treated and untreated units, or $\widehat{ATE} = \overline{Y_t}(1) - \overline{Y_c}(0)$.

This estimate of the ATE, however, does not permit inferences about ITEs. At best, and in the absence of any knowledge about treatment effect covariates such as moderators and mediators as well as heterogeneity in their individual-specific realizations, the ATE estimate can figure as the expectation of the ITE for an individual randomly drawn from the experimental population. But as soon as there is (suspected) heterogeneity among treatment-effect covariate realizations and consequently ITEs, this estimate will not be

[4] The distinction between confounders and moderators/mediators being that confounders influence the outcome variable independently of treatment whereas moderators/mediators influence the outcome by affecting the causal pathway(s) connecting treatment and outcome.

precise, so accurate inferences about ITEs are largely precluded and information on heterogeneity cannot be recovered from $\widehat{ATE}$.[5]

This has significant bearing on the *transferability* of trial results, i.e. the extent to which the ATE from a study population A can be expected to be replicated in some other population B. Two jointly sufficient conditions for the transferability of trial results to some out-of-sample target are first, that the treatment variable plays the same causal role in the production of the outcome in the target as it does in the experimental population, i.e. that the mechanisms in both populations are sufficiently similar with respect to the causal claim to be extrapolated. The second condition is that the distribution of treatment effect covariates in the target is the same in both populations (see e.g. Cartwright and Marcellesi 2015 for similar conditions).[6] So the transferability of experimental results to targets hinges not only on sufficient similarity in mechanisms between populations but also on whether there is heterogeneity effected by differences in treatment-effect covariates as well as how such covariates such as moderators and mediators are distributed among agents in the populations of interest. This problem has received attention from a variety of econometricians, methodologists, philosophers of science and EBP proponents (e.g. Hotz-Imbens and Mortimer 2005, Duflo, Glennerster and Kremer 2008; Imbens and Wooldridge 2009; Bareinboim and Pearl 2013; Cartwright and Marcellesi 2015).

However, heterogeneity does not only affect the transferability of trial results. It also creates a second challenge for EBP. The challenge is that in the absence of information on heterogeneity, RCTs are not suitable for informing *any* policy formation process that is concerned with the *distributive consequences* of policy (cf. Manski 2000). More specifically, policy-makers are often interested in knowing not only whether an intervention is effective on average but also in how effective the intervention will be for specific types of agents, how heterogeneous treatment effects are distributed among agents, with respect to which observable baseline characteristics, whether heterogeneity obtains in magnitude or also in sign, etc.

This information is crucial particularly in those cases where it is reasonable to suspect that at least some agents may respond negatively to an intervention, even though the ATE might be positive. In these scenarios, several pertinent distributive concerns arise, e.g. is it at all permissible to implement policy that will render some agents worse off? If so, how

[5] While this difference-in-means estimation yields, without strong assumptions, an unbiased estimate of the sample ATE, and under somewhat stronger assumptions of the population ATE, it takes substantive assumptions about distributions of ITEs to estimate even the sample variance of the ATE (although this estimate can be bounded by inspection of the treatment and control mean variances).

[6] Necessary conditions might be weaker, cf. Bareinboim and Pearl (2013)

should we adjudicate between the negative welfare consequences for these agents and the net effectiveness of the intervention? What are the thresholds of proportionality that we should use to decide whether welfare benefits on the part of some outweigh welfare losses on the part of others? Can the policy be targeted so that it predominantly affects those who will benefit from the intervention? And so forth.

As these stylized concerns suggest, policy-makers may be interested in pursuing a variety of different distributive values. Yet, in order to pursue these values rigorously, in the sense that they have good reasons to believe that an intervention will promote them, policy-makers require information on treatment effect heterogeneity, i.e. whether there is heterogeneity at all and how heterogeneous treatment effects are distributed with respect to agents' observable characteristics. As I have argued above, RCTs do not provide such information on their own.

Yet, this does not mean that EBP methodology is at a complete loss in this regard, as EBP proponents may be keen to point out that one way to address this problem is to perform so-called *subgroup analyses*. However, I argue below that performing such analyses comes at the expense of sacrificing several key EBP epistemic values and that this creates a tradeoff between the epistemic values central to EBP and the pursuit of moral and political values such as equality and priority for the worst off.

3.2 Subgroup Analysis as a Remedy for Informing about Heterogeneity

Following Duflo, Glennerster and Kremer (2008), subgroup analyses partition experimental populations into subgroups according to observable characteristics such as age, sex, ethnicity, prior education etc. They then typically further partition subgroups into different categories or strata, for instance age groups. Given this stratification, a difference-in-means estimation can be run on the partitioned data to obtain conditional, subgroup-specific ATEs (CATEs). An alternative to this stratification approach that is applied predominantly when investigating binary and categorical variables, is to run so-called *meta-regressions*, where potentially interesting treatment-effect covariates are modeled as interaction terms with treatment in a standard regression framework. In doing so, it is possible to obtain information on significant interaction effects between observables and treatment that may be taken as evidence for the involvement of the respective treatment effect covariates as moderators or mediators.

Even so, while subgroup analyses seem to offer at least tentative information about heterogeneity, they are also subject to several pertinent methodological concerns. Let me expand on two particularly pressing concerns and explain how they bear on the realization of EBP epistemic values.

First, the information that meta-regressions can generate is purely correlational in nature, and hence subject to standard concerns about endogeneity and consequent bias. For instance, statistically significant parameter estimates on treatment effect heterogeneity of microfinance programs with respect to differences in prior business ownership do not permit the straightforward interpretation that prior business ownership is a causally relevant treatment effect covariate.

This is because the significance of the estimate may be attributable to common-causes, e.g. because business ownership is highly correlated with business education, and it is business education that is causally relevant for the production of microfinance outcomes, but prior business ownership in the absence of business education may not contribute at all to outcomes of interest.[7] In this case, if business education is not included in the regression, our estimates of individual-level heterogeneity with respect to prior business ownership will be biased.

More generally, parameter estimates for treatment effect covariates will invariably remain subject to such concerns about bias unless we can entertain the relatively strong assumption that the regressors are uncorrelated with the error term of the meta-regression (see e.g. Pearl 2014). However, it is precisely such assumptions, which are necessary for unbiased identification in regression contexts, that EBP proponents are typically keen to avoid and that are expressly dismissed in the methodological tenets that emphasize randomization as the key strategy to avoid questionable identification assumptions.

Randomization at the treatment stage does not alleviate these concerns either, because treatment effect moderators are not necessarily randomly distributed among agents who, with respect to one subgroup characteristic, may *systematically* differ on several other relevant and collinear or interacted covariates at once. This means that obtaining *unbiased estimates* and straightforward *causal conclusions* about the role of covariates as treatment moderators is typically precluded, threatening at least two EBP epistemic values at once.

A second worry about subgroup analyses concerns the *precision* of effect estimates and *statistical power*. In short, the more subgroups one specifies, the higher the probability of obtaining spurious results. For typical significance levels at $p < 0.05$ even a moderate number of subgroups, strata partitions and corresponding hypothesis tests will render the

[7] For instance, prior business ownership in the absence of business education can be exhibited by agents who have previously pursued unprofitable business plans and may continue to do so in the future. Thus the unbiased parameter estimate for business ownership is likely to be substantially smaller than the estimate for business education. To permit unbiased estimation of interaction terms, one would at least need to induce additional exogenous variation in the covariates of interest. But this would require significantly different trials designs with multiple, parallel interventions on treatment as well as covariate realizations (see e.g. Imai et al. 2013). While such designs are in principle feasible, they also raise issues with precision and statistical power.

occurrence of spurious results exceedingly likely. At the very least, suitable statistical corrections for multiple hypothesis testing are in order to remedy the consequences of multiple testing for the prevalence of false positives. Yet, while recommended by some EBP proponents (e.g. Duflo, Glennerster and Kremer 2008, 65), this is rarely carried out in practice (cf. Fink et al. 2014, 47). Moreover, to alleviate concerns about insufficient statistical power and precision, sample sizes may need to be expanded for subgroup analyses to be sufficiently informative. For instance, in order to detect a heterogeneity signal of the same magnitude as the ATE and with the same precision as the ATE estimate, a difference-in-means estimation on just one subgroup partitioned into two strata requires a fourfold expansion of the original sample size (Varadhan and Seeger 2013, 38). Yet, subgroup-specific effects are often significantly smaller than ATEs, which requires much greater expansions of sample size to maintain sufficient power.

These and other, related concerns severely limit the extent to which subgroup analyses can inform about treatment effect heterogeneity. At most, and in line with standard recommendations (e.g. Varadhan and Seeger 2013), subgroup findings should be considered *exploratory* in the sense that they may prompt additional investigations such as novel trials on subgroups of interest, but are insufficient to warrant definitive conclusions about heterogeneity by themselves.

However, while conducting novel trials on potentially vulnerable subgroups appears to be a viable strategy to address some of the above concerns, this requires prior identification of the relevant subgroups. Unfortunately, we are rarely in the epistemically fortunate position to know which individuals are most likely to incur welfare losses in advance, since that depends on knowing what the causally relevant treatment effect covariates are, how they affect the outcomes of interest as well as which agents exhibit beneficial vs. harmful realizations of such covariates. So precise information on heterogeneity is still required even if we are willing to conduct subsequent trials on vulnerable subgroups.

The extant EBP literature has only recently started to address treatment effect heterogeneity issues. Yet, even though there are several recent social policy and development studies that perform at least tentative and exploratory heterogeneity analyses, they frequently fail to address one or more of the concerns outlined above (see e.g. Fink et al. 2014) or tend to focus on *between-trial* heterogeneity, which is a related but conceptually distinct issue from the *within-trial* and *between-subject* heterogeneity that I consider here.

Let me expand on how these epistemic challenges for informing about heterogeneity create a trade-off between epistemic and moral values and how this trade-off challenges both value-freedom and neutrality in EBP.

4. A Trade-off Between Epistemic and Non-Epistemic Values

The trade-off between epistemic and non-epistemic values that I want to highlight is a result of the differential usefulness of EBP research outcomes for the pursuit of different kinds of moral values, i.e. broadly utilitarian and non-utilitarian values respectively.

Standard EBP methods such as RCTs, Regression Discontinuity Designs and IV identification strategies are in general capable of generating outputs that are sufficient for the pursuit of standard utilitarian values, i.e. those that are concerned with the increase or maximization of aggregate or average welfare. This is because the distribution of individual-specific contributions to aggregate welfare outcomes is not a primary concern for increasing aggregate or average welfare, so information on heterogeneity is not necessary for the pursuit of these values.[8]

Yet, such information on heterogeneity is necessary for the pursuit of *any* moral and political value that is sensitive to how aggregate outcomes are realized. For instance, the pursuit of broadly egalitarian or prioritarian values requires at least information on the initial distribution of welfare among agents as well as information on the changes to this distribution brought about by the intervention at issue. Yet, as I have argued above, such information on treatment effect heterogeneity cannot be provided by RCTs alone. At the very least, subgroup analyses need to be carried out in order to permit at least tentative conclusions about heterogeneity. Moreover, methods such as Causal Bayes Net Analysis, Qualitative Comparative Analysis, Process Tracing and Machine Learning may present potentially superior alternatives for the identification of causally relevant treatment effect covariates that generate heterogeneity. However, such techniques are rarely acknowledged or mentioned in the standard manuals circulating in the EBP literature (e.g. Angrist and Pischke 2009), and even if they were, these methods are often neither straightforwardly compatible with the identification strategies that EBP practitioners typically pursue nor with the evidence ranking schemes that EBP methodologists subscribe to.

This licenses two conclusions. First, EBP methodology presently favors the production and use of evidence suitable for the pursuit of utilitarian values, i.e. those that focus on increasing or maximizing average or aggregate welfare. Second, EBP methodology presently fails to adequately promote or even hinders the production of high-quality evidence on heterogeneity that is necessary for the pursuit of many non-utilitarian values. As a consequence, standard EBP methodology renders the pursuit of distributive values such as egalitarian or prioritarian ones relatively more difficult or infeasible.

[8] It might still be helpful, since welfare *maximization* is easier to accomplish when we have information that helps pick out those individuals who will likely benefit most from some intervention; granted that interventions can be targeted to affect only such individuals.

This generates a trade-off between the epistemic values central to EBP and the moral and political values that policy-makers are in a position to pursue effectively on grounds of EBP evidence. More specifically, whenever the pursuit of moral and political values requires information on distributive consequences of policy, standard EBP evidence fails to provide the required information. Conversely, whenever evidence of the kind required to inform about distributive consequences of policy shall be produced, this requires at least some sacrifice of basic EBP epistemic values. More specifically, whenever EBP methodology and methods are changed in order to generate information on heterogeneity, e.g. by means of subgroup analyses, this comes at the expense of sacrificing at least three crucial EBP epistemic values at once, i.e. the *unbiasedness* and *precision* of effect estimates, as well as the *ability to obtain causal conclusions*. Maintaining these values, on the other hand, comes at the expense of sacrificing the informativeness of EBP research outputs about the distributive consequences of policy.[9]

Let me expand on what this trade-off implies for value-freedom and neutrality in EBP. First, if the value-free ideal underlying the EBP paradigm is to say that non-epistemic values are generally not and should not be involved in shaping the conduct and outcomes of EBP research internally, then the desirability of this ideal is challenged. The reason is that moral and political values are at least involved to the extent that without suitable changes to EBP methodology, the pursuit of non-utilitarian values is inhibited. If this situation should be remedied, then this requires changes to methodology that privilege or prioritize the production of evidence on heterogeneity. However, and this is the crucial point, these changes will be *effected by moral values*, since it is the pursuit of moral values that motivates the requisite changes to methodology. To the extent that these changes to methodology are justifiable and justified, this means that value-freedom in EBP is not a desirable ideal, even at internal stages such as method choice and model specification.

Value-neutrality is challenged as well. It assumes that once the desirability of some social outcome is agreed upon, evidence speaking in favor of the effectiveness of some intervention in realizing this outcome at most figures in conditionally normative policy recommendations.

[9] This point may appear similar to Helen Longino's who argues that several traditional epistemic values are not purely epistemic and "[…] that their use in certain contexts of scientific judgment imports significant socio-political values into those contexts" (Longino 1996:54). However, my point is weaker than Longino's in the sense that it should appeal even to those who insist on the purely epistemic character of values such as unbiasedness, precision, and the ability to obtain causal conclusions. Specifically, I do not argue that these values fail to be purely epistemic as they exhibit a demonstrably political (or moral) valence (ibid.). Instead, even if we grant that these values are purely epistemic, their pursuit may still have important ramifications for the extent to which the pursuit of other, moral values is facilitated or inhibited.

Yet, inferences about policy effectiveness are typically grounded in information about ATEs and as such do not accommodate information on distributive consequences. So this way of operationalizing what it means for a program to be effective brackets concerns about heterogeneity. As it stands, an *effective* program is considered a good program to the extent that the outcome of interest tracks a relevant moral or societal good. However, even if this good is uncontroversial in itself, effectiveness still only means effectiveness on average, not some effectiveness for everyone, or sufficient effectiveness for the worst-off, or equal effectiveness for all policy subjects.

To maintain neutrality with respect to distributive values it is not enough to agree on the desirability of social outcomes *as such*. It is also necessary to agree upon the *ways in which* these outcomes may be realized, since a given change in aggregate outcomes can usually be achieved in various ways, each of which may have dramatically different distributive consequences for target populations, some of which may be more or less desirable *in themselves*. This issue is masked when broadly utilitarian values are pursued, but becomes apparent when distributive consequences matter; as is the case for the pursuit of egalitarian and prioritarian values. So if we care about differences between agents and about absolute and relative changes in outcome distributions, then *effectiveness* as standardly construed in EBP is not informative about the moral permissibility or desirability of policy and might be misleading about what *effective* programs are ultimately able to do for us, given the specific moral and political values that we pursue.

So at present, it seems that the dissemination of EBP research is premised on the implicit value presupposition that the relevant magnitude for deciding which policy to implement is its effectiveness in terms of average treatment effects. And this fails to be value neutral in the envisioned sense because it assumes that average effectiveness is the proper target of interest rather than delegating the question of whether it is, to policy makers and other agents to settle. In a nutshell, in order to maintain a traditional ideal of value-neutrality, additional value presuppositions such as the above must be made explicit for EBP policy recommendations to remain value-neutral in the envisioned sense.

5. Conclusion

I have argued that there exists a trade-off relation between key EBP epistemic values and non-epistemic values that are sensitive to distributive consequences of policy, e.g. equality and priority for the worst-off. This trade-off obtains because the outputs afforded by standard EBP methods are differentially useful for the pursuit of different moral and political values. I have argued that this trade-off challenges ideals of value-freedom and neutrality in the EBP paradigm. This may be taken as starting point to reconsider some of the standard epistemic value presuppositions entertained in EBP as well as for refining

EBP methodology in ways that enable and facilitate the pursuit of a wider range of moral and political values.

References

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. "Mostly harmless econometrics." Princeton: Princeton University Press.

Banerjee, Abhijit, Emily Breza, Esther Duflo, and Cynthia Kinnan. 2015. "Do Credit Constraints Limit Entrepreneurship? Heterogeneity in the Returns to Microfinance". Working paper.

Bareinboim, Elias, and Judea Pearl. 2013. "A General Algorithm for Deciding Transportability of Experimental Results." Journal of Causal Inference. 1: 107-134.

Cartwright, Nancy, Alexandre Marcellesi. 2015. "EBP : Where Rigor Matters." In Foundations and Methods from Mathematics to Neuroscience : Essays Inspired by Patrick Suppes, ed. Colleen E. Crangle, Adolfo García de la Sienra, Helen E. Longino, Stanford: CSLI Publications.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." In Handbook of Development Economics, Vol. 4, ed. Paul T. Schultz and John Strauss. Amsterdam and New York: North Holland.

Fink, Günther, Margaret McConnell, and Sebastian Vollmer. 2014. "Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures." Journal of Development Effectiveness. 6:44-57.

Hausman, Daniel, and Michael S. McPherson. 1996. "How Could Ethics Matter to Economics?" In Economic Analysis and Moral Philosophy, Hausman and McPherson, Appendix. Cambridge: Cambridge University Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." Journal of the American Statistical Association. 81:945-970.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." Journal of Econometrics. 125:241–270.

Imai, Kosuke, Dusting Tingley, and Teppei Yamamoto. 2013. "Experimental designs for identifying causal mechanisms." Journal of the Royal Statistical Society A, 176 Part 1:5-51.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." Journal of Economic Literature. 47:5-86.

Kuhn, Thomas S. 1977. *The Essential Tension*. Chicago, IL: University of Chicago Press.

Longino, Helen. 1996. *Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy, in Feminism, Science, and the Philosophy of Science*, ed. Lynn Hankinson Nelson and Jack Nelson, 39-58. Dodrecht: Kluwer.

Manski, Charles F. 2000. "Identification problems and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice." Journal of Econometrics. 95(2):415–442.

Oxman, Andy D., and Gordon H. Guyatt. 1992. "A Consumer's Guide to Subgroup Analyses." Annals of Internal Medicine. 116:78–84.

Pearl, Judea. 2014. "Reply to Commentary by Imai, Keele, Tingley and Yamamoto Concerning Causal Mediation Analysis." Psychological Methods. 19(4):488-492.

Rubin, Donald. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." Journal of Education Psychology. 66:688-701.

Varadhan, Ravi, and John D. Seeger. 2013. "Estimation and Reporting of Heterogeneity of Treatment Effects." In Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide, ed. Patricia Velentgas, Nancy A. Dreyer , Parivash Nourjah, Scott R. Smith, Marion M. Torchia, 35-44. Rockville, MD: Agency for Healthcare Research and Quality.

Synthetic Biology and the Search for Alternative Genetic Systems: Taking How-Possibly Models Seriously

Rami Koskinen, *University of Helsinki*

Contributed paper, PSA2016

Word count: 4838

Abstract

How-possibly models are usually treated as some kind of second-rate theoretical tools. They may be indispensable in the early stages of theorizing, but do not constitute the main aim of modeling, namely, the discovering of a one true mechanism responsible for the phenomenon under study. I argue that this prevailing picture does not do justice to the synthetic strategy that is commonly used in the engineering sciences. In synthetic biology, how-possibly models are not something to be eliminated by a more detailed analysis, but rather design hypotheses for a field whose ultimate goal is to build novel biological systems.

1. Introduction

According to the most influential contemporary account of explanation in the philosophy of biology, namely, that of mechanistic explanation, the aim of the life sciences is seen as the discovering and modeling of mechanisms that "produce, underlie, or maintain a phenomenon" that is being studied (Craver and Darden 2013, 15). The mechanistic strategy of modeling is typically conceptualized as proceeding by somehow constraining a space of possible mechanisms for a given phenomenon or function. According to Craver (2007, 31), the space of possible mechanisms contains all the mechanisms that could possibly explain a phenomenon. By explicating a particular point in this space, scientists construct a *how-possibly explanation* or *model*. Furthermore, it is often assumed that there is just one true or correct mechanism, the details of which are ideally captured in a finished *how-actually model*. Intermediate between these two extremes are *how-plausibly models*, which form a more tightly constrained subset of how-possibly models, but still lack the full empirical support of a how-actually model. (Craver and Darden 2013, 34–35.)

According to the prevailing picture of biological modeling strategy, a successful search for a mechanistic explanation should converge on one single mechanistic model candidate, and divergent how-possibly models that differ in their mechanistic details should be discarded as superfluous and scientifically incorrect. According to Craver (2007, 131), "Distinguishing good explanations from bad requires that one distinguishes real components from fictional posits. The most dramatic examples of fictional posits include animal spirits, entelechies, and souls, but fictitious entities can be far more mundane than these". He concludes that many how-possibly mechanisms "require parts (and activities)

that *do not exist*" (Craver 2007, 131, my emphasis). Because modern-day scientists want their models to work, and in particular, do not want to commit themselves to any kind of spooky non-existent entities, how-possibly models are usually considered as something that should be eliminated as quickly as possible when conducting serious research. In contemporary philosophy of science, how-possibly models are often treated as some kind of second-rate explanations or theoretical tools (e.g., Rosenberg 2006, 45; see also Craver 2006, 361, 2007, 112).

In this paper I will argue that this current view concerning the role of how-possibly models is very narrow. More precisely, it may be a good approximation in the context of scientific analysis of natural systems where research advances through the methods of decomposition and localization (Bechtel and Richardson 1993/2010). However, this does not preempt all the goals of biological investigation. The idea of starting from a range of possible models and then working towards one or a very limited number of how-actually models seems to make much sense when one considers the general purpose of biological investigation. For example, given that one of the main aims of science is to provide manageable generalizations that unify phenomena as much as possible, focusing on how things actually work is a neat idea and surely a good starting point! Especially, it is much more manageable to model complex high level input–output phenomena when they can be cashed out in terms of a few select mechanisms that are already familiar. Another reason that is especially prominent in the life sciences is the ability to effectively intervene on various target systems for medical purposes (see Craver 2007). Why would scientists

bother wasting their time with mere how-possibly models that do not provide good access to actual phenomena, not to mention ways to effectively intervene on them?

However, although it is often the case that scientists are interested in some well-defined actual target system, it is also true that a lot of times the target of investigation is some more abstract feature of the living world that might require studying objects that, strictly speaking, do not exist, at least at the moment of investigation (Dawkins 1986; Dennett 1995, 102–103). I hold that the same is true also in the context of the synthetic strategy that is commonly used in the engineering sciences. In the field of synthetic biology, researchers use how-possibly models to study what may be called potential biological systems. I argue that in the hands of bioengineers, abstract how-possibly models are not something to be eliminated by a more detailed analysis, but rather design hypotheses for a field whose ultimate goal is to build novel biological systems and "re-wire" existing ones. I explicate this role further by providing an example from the study of alternative genetic systems by synthetic biologist Steven Benner and his group. The case will highlight how the method of synthesis, even when it fails, provides an effective way to limit the space of possible models for biological mechanisms. This has effects for the study of potential and actual natural systems alike.

2. From Actual to Potential Biological Systems

It is often said that one important thing about mechanistic understanding is the ability to answer "what-if-things-had-been-different" questions (e.g., Craver 2006, 358, following

Woodward 2003). This is certainly true in the sense that, ideally, when a mechanism is fully understood (i.e., our best model of it does not contain any black boxes left to open) we are able to reliably predict its output for a range of input and parameter values and even manipulate its functioning. Knowing how an actual mechanism operates as accurately as possible gives us more effective ways to handle typical contrafactual questions that arise in science (cf. Craver 2006). However, this kind of access to full mechanistic details of actual target systems forms only one part of contrafactual reasoning that is of interest to scientists. Sometimes, especially when dealing with some more theoretical issues, scientists who ask "what-if-things-had-been-different" questions are not in fact inquiring how accurately we understand the parts and workings of some actual mechanism. Rather, I suggest, they might be wondering whether the mechanism (or the system in general) itself could, or could have been, different. It is in this way that, instead of being just an eliminable scaffold on the way towards a how-actually model, a how-possibly model can become the main object of inquiry in its own right.

Taking how-possibly models of biological systems seriously in the above sense might mean two things. First, it might simply mean taking seriously the general strategy of "turning the tables around", that is, focusing research on what is possible instead of actual in the biological world. This is akin to an exploration into the dark where rather few things limit the search space. Second, it might mean that one is committed to the study of some particular how-possibly model or set of models for a phenomenon for which there already is a how-actually model. The second version has the nice advantage that we already have an existential proof that *that* phenomenon or function is indeed realizable at all. We can

then investigate whether it can be achieved by means of some alternative mechanism; the strategy is essentially *contrastive* in nature. Indeed, this is something that is done in many quarters of biological engineering and especially in the field of synthetic biology.

As in the case of more traditional life scientific research, mechanistic understanding and modeling of biological systems is at the heart of synthetic biology. In a sense, synthetic biology can be seen as taking them even further. The field is often characterized by a strive to build novel biological systems (Elowitz and Lim 2010). Because this requires an excessive ability to manipulate existing biological mechanisms, synthetic biology is often portrayed as the ultimate test for our mechanistic understanding of the living world in general (Endy 2005; Elowitz and Lim 2010).[1] However, at the same time these bioengineers are also testing completely new waters by expanding biological understanding over and above naturally occurring systems. Some of the synthetic systems, like artificial genetic circuits, that have been built can be seen both as new biological objects in their own right and as certain kind of concrete, but theoretical models of what kind of design principles are biologically feasible (Knuuttila and Loettgers 2013). Although the study of these potential biological systems is targeted at certain very specific types of how-possibly models of biological systems, they can be seen as enriching our understanding of biology in terms that go beyond mere engineering feats (Morange 2009).

1 Craver and Darden (2013, 92–94) also mention the importance of engineering or "build it" test as an effective way to further refine scientific understanding of biological mechanisms.

For example, it has been suggested that synthetic design methods might be able to prove valuable information about the nature and limits of the evolution of gene regulation:

> [An important problem in evolutionary biology is] why the genetic network architectures we observe in Nature have evolved to solve a particular problem an organism faces in its environment. This challenge is often complemented by the question of which selective forces (i.e. environmental or cellular conditions) have shaped the biological systems we observe in modern organisms. The null hypothesis is simply that a particular architecture has arisen by non-selective forces and that multiple architectures would be sufficient to achieve the biological functionality observed. (Bayer 2010, R775.)

In normal evolutionary research, these kinds of questions are often difficult to evaluate because many specific functions are found only in a very limited number of systems or model organisms; the relevant sample might also be biased by historical contingency. Modeling the mechanistic details of the actual system in ever greater detail does not seem to be of much help here. However, thanks to synthetic biology and other forms of biological engineering, evolutionary studies are no more necessarily restricted by the availability of naturally occurring systems: “The construction of synthetic versions of natural circuits is a powerful way to interrogate questions of ‘why’ in biology” (Bayer 2010, R775). Endy also defines one of the main goals of synthetic biology as follows: “[…] synthetic biology provides an opportunity to test the hypothesis that the genomes encoding natural biological systems can be ‘re-written’, producing engineered surrogates

that might usefully supplant some natural biological systems" (Endy 2005, 449; see also Sprinzak and Elowitz 2005).

The re-design strategy depicted here limits its focus on systems that differ in their underlying mechanistic architecture, but that are nevertheless capable of realizing the same higher level function. It is reminiscent of the situation that philosophers often call by the name "multiple realizability". Because how-possibly models are often presented in exactly this kind of situation where they are in a sense explanatory rivals for one and the same phenomenon, it is easy to see how they fit into the conceptual scheme of "biological re-writers". One of the most compelling examples of this kind of research comes from the study of artificial genetic systems that can be regarded as functional alternatives for our natural DNA. It is there that various how-possibly models, on top of their more traditional explanatory purport, seem to have the role of explicit design hypotheses.

3. Alternative Alphabets for Life's Code

Why is the genetic code based on the DNA molecule? Is it a functional necessity, or just a historical accident? Because the sample size of life on Earth is one, there is no straightforward empirical way to investigate this issue. In his famous booklet *What is Life?* the physicist Erwin Schrödinger (1944) originally proposed an inspirational how-possibly model for genetic material in which genes were hypothesized as consisting of some kind of aperiodic crystals. This was nine years before Watson and Crick's discovery of the structure of the DNA molecule. Because of their groundbreaking work, we, of course, now

have an excellent how-actually model for the implementation of the genetic material. However, as successful as their model has turned out to be, it does not really answer all the why-questions that can be raised regarding the material nature of the genetic code. To answer these questions would require contrasting DNA with some other plausible how-possibly models for genetic material and hoping for some principled clue as to why nature has opted for this particular solution.

Beginning already in the late 1980's (Switzer, Moroney, and Benner 1989), synthetic biologist and chemist Steven Benner has been studying what can be called artificial genetic systems. These are chemical structures that are supposed to have the essential functional features of a genetic code, but that are nevertheless different from the structural design of familiar DNA and RNA molecules. According to Benner:

> In a version popular today in some engineering communities, [synthetic biology] seeks to use *natural* parts of biological systems (such as DNA fragments and protein "biobricks") to create assemblies that do things that are *not* done by natural biology (such as digital computation or manufacture of a speciality chemical). […] Among chemists, "synthetic biology" means the opposite. Chemist's "synthetic biology" seeks to use *unnatural* molecular parts to do things that are done by natural biology. Chemists believe that if they can reproduce biological behavior *without* making an exact molecular replica of a natural living system, then they have demonstrated an understanding of the intimate connection between molecular structure and biological

> behavior. If taken to its limit, this synthesis would provide a chemical understanding of life. (Benner, Yang, and Chen 2011, 372; emphasis in original.)

For the last 25 years Benner has done just that. That is, he has studied a wide range of different chemical systems that could potentially be used to fulfill the same role as DNA/RNA in naturally evolved organisms. Although no such system still exists, researchers have managed to construct many interesting variants that have at least some of the features required of a code of life, and new exciting results are frequently being reported from scientists working at the junction between synthetic biology and chemistry (see, e.g., Malyshev at al. 2014; Marlière et al. 2011; Thyer and Ellefson 2014).

Although these studies are limited and far from conclusive, they nevertheless provide reasons to believe that DNA might not be the kind of necessary ingredient that some take it to be as the only thing capable of turning inanimate matter into living, reproducing, and evolving systems. In the language of Craver and Darden (2013, 69), they give us good reasons to suppose that alternative genetic systems remain a *live possibility*. To make the continued hegemony of DNA as the biochemical medium of choice seem even less secure, Benner noted in an interview published in *Nature* in November 2012 that “The first thing you realize is that [DNA] is a stupid design”, further insisting that, “If you were a chemist setting out to design this thing, you would not do it this way at all.” (Kwok 2012, 516.)

However, it is one thing to criticize the structural design of DNA, and another to show that any other chemical solution would be able to perform the same functions. Because all known organisms have their genetic code based on DNA/RNA, the possibility of

"alternative genetic alphabets" requires strong empirical proof (Thyer and Ellefson 2014, 291). One would expect there to be good chemical and evolutionary reasons for DNA to be the medium for genetic information. Although it is nowadays recognized that biological solutions are not always optimal in the strong sense, they are nevertheless often extremely robust and surprisingly efficient (see Wagner 2005). Because so much complex, evolutionarily successful life is based, at the bottom level, on the structural features of DNA, it simply cannot be *that* inadequate as a molecule. However, it is also because of this most intense of dependencies that it is actually very difficult to make many far reaching biological conclusions about the nature of DNA; it is a deeply generatively entrenched fact about the living world (Wimsatt 2007, 135–136).

To understand the requirements for an adequate genetic code, I will first have to examine the definition of a living system that Benner and others advocate. This is a working definition, which means that it is open for revisions. It nevertheless captures many of the features that biologists take to be essential for living systems. In his work, Benner follows the so-called "NASA definition" of life as a self-sustaining chemical system capable of Darwinian evolution (Benner, Yang, and Chen 2011, 375). Although this definition leaves many important facts up for further refinement, it nevertheless already makes some empirical bets by, for example, ruling out genuinely Lamarckian systems. As Benner himself notes, we do not have any reasons to believe that even Lamarckian systems would be strictly impossible (Benner, Yang, and Chen 2011, 375). However, we have to start from somewhere, and at least we have many empirical examples of Darwinian life, not to mention a particularly successful evolutionary framework that unifies these findings. In a

sense, synthetic biologists can take the testing of this theory even further by trying to come up with some other kind of chemical systems that can be subsumed under it.

Benner's abstract model of a genetic system has three features: (1) the ability to carry biological information, (2) the ability to transmit biological information, and (3) the ability to support Darwinian evolution (based on Benner, Yang, and Chen 2011). In practice, the above list can be thought to encompass also some implicit auxiliary assumptions à la Duhem and Quine. Examples of these could be some kind of linear arrangement of the code, or the overall chemical and thermodynamic stability of its structure (see Szathmáry 2003). These general features can also be broken down into smaller mechanisms or causal role functions that make them physically feasible. For example, the encoding of biological information is often taken to require some kind of chemical specificity, like bonding, lock-and-key complementarity, and so on. This brings the whole enterprise closer to the how-plausibly end of all conceivable possibilities.

Although the above list seems quite simple, it contains an implicit tension that makes it more difficult to achieve all of the requirements simultaneously. It is obvious that without the first requirement, we would simply have no code at all—genetic or other kind. However, the mere ability to store information is not that interesting property in itself. The code must also be able to transmit information from one system to another. It is only after this step is fulfilled that we can actually speak about inheritance. Taken in isolation, the requirements (1) and (2) suggest that the more accurate the functions in question are, the better the medium is in realizing the code. In a sense this is true. A system that transmits its

information content so poorly that descendant systems hardly resemble their parent systems would clearly not be able to support Darwinian life. However, what the requirement (3) implicitly insists is that although the copying process should be reliable, it should not be completely certain. Otherwise no variation is ever going to accumulate, and the system can only produce an endless army of genetically identical clones. Again, the space of possible models for genetic material is in this way constrained before any considerations about the physical medium has taken place.

According to Benner, many synthetic biologists' original expectation was that the best place to start changing the chemical basis of the genetic code was the sugar backbone of natural DNA molecules. This is because the informational specificity of the genetic code is often thought to lie in the highly specific complementary base pairings between the nucleobases A-T and C-G. The backbone was believed to be just a contingent structure, a kind of molecular "scaffold", whose purpose is to support the real sources of information. (Benner, Yang, and Chen 2011, 376.) However, as it turned out after numerous trials, Benner and his team were unable to achieve functionally stable molecules by changing the sugar backbone. For example, backbones made of glycerol units turned out be too flexible and the whole structure broke down in normal temperatures; the nucleobase bindings alone were not strong enough to hold the structure together. (Benner, Yang, and Chen 2011, 377.)

In addition to sugars, the DNA backbone features also phosphates. Similar results were attained by Benner and his team when they tried to change the phosphates as did in the case of the sugars. For example, when the phosphates were replaced by synthetic

oligosulfones, the structure tended to fold onto itself. (Benner, Yang, and Chen 2011, 378.) This was bad news, because folded structures might not be specific enough to ensure faithful pairing. Moreover, if the structure is stripped off of its repeating charges that are manifest in the phosphates, it might hamper its mutability; remember that the ability to evolve is one of the functional requirements of life that Benner advocates in his working model. (Benner, Yang, and Chen 2011, 379–380.) Thus, the sugar and phosphate backbone with its repeating charges seemed to be a necessary feature of a biologically plausible genetic system capable of Darwinian evolution.[2]

This meant that in order to achieve the grand goal of an alternative genetic system, the changes must be made to the nucleobases. After trying out numerous working hypotheses or possible models for a genetic code that is based on alternative or "unnatural" alphabets, Benner and his team finally arrived at the six letter alphabet (A, T, C, G, P, Z). Using methods from modern biological engineering, like polymerase chain reaction, Benner and his team were able to add the bases P and Z to a system based on the natural (A, T, C, G) alphabet. These bases were selected because bonding between them has experimentally been shown to be very strong and specific. Furthermore, unlike in the case of alternative backbones, the new bases *can* support Darwinian evolution: In the case of the (A, T, C, G, P, Z) alphabet, the new bases have been shown to be mutable to natural bases. Also, and

2 In the case of some *xeno nucleic acids*, researchers have been able to change the sugar backbone of DNA molecules. However, it is not clear whether these systems can support life. See Schmidt (2010).

perhaps even more interestingly, C and G could mutate to give P and Z. (Benner, Yang and, Chen 2011, 384.)

Benner's alphabet has some very interesting biological properties. First, it is a mix of both natural and unnatural biochemical, or genetic, "letters". Second, its cardinality differs from that of the natural DNA/RNA alphabet. These two features make it possible to use the new alphabet to study both disparate *and* expanded genetic alphabets at the same time—a double win. I maintain that these features also make it an interesting case to study various how-possibly models of genetic systems.

Do the incorporated bases P and Z radically differ from the molecular structure of those of the natural nucleobases? This is somewhat debatable. It is true that the molecular mechanisms of pairing between them resemble those of A-T and C-G. However, they are still structurally different molecules. It was certainly far from clear that these bases could be inserted successfully into a system of natural alphabet. This is especially so because they tend to change the "dynamics" of the whole structure. The more different types of parts that can possibly interfere with each other there are, the more likely it is that the whole system will fail to be able to perform its functions properly; the basic parts of a mechanism often seriously constrain its space of possible models (cf. Craver and Darden 2013: 105). Also, the informational change that is brought by the new nucleobases can be as interesting as the change in structural features. With alphabets of different cardinality, it is possible to test the idea whether the functions of living systems can be coded in ways that differ from the familiar four-letter system. Although previous models had been

susceptible about this (Szathmáry 2003), Benner's (A, T, C, G, P, Z) alphabet is one of the first experimental results to give compelling reasons to believe that they do.

4. Conclusions

Contrary to traditional life-scientific practice, the engineering or "synthetic" strategy of fields like synthetic biology can be used to explore the space of biological possibilities by taking established how-actually models for biological systems as a starting point, and then working towards realizing alternative and contrasting how-possibly models. If it succeeds, it opens up new exciting possibilities. For example, a fully functional alternative genetic alphabet could work as a genetic "firewall" between engineered and naturally evolved organisms, providing an effective biosafety tool (Schmidt 2010). However, even if it does not, something new is still being learned about the nature of actual systems. Both situations can have benefits for basic science.

Besides Benner's group, many others have also been working on synthetic genetic systems and alternative alphabets. For example, Marlière et al. (2011) report a successful incorporation of a new nucleobase 5-chlorouracil into a laboratory strain of *E. coli*, while Malyshev et al. (2014) produced similar experiments with the pair d5SICSTP-dNaMTP. What is remarkable is that both cases exemplified robust functionality with no obvious biological pitfalls. It might be that some of the structural features of DNA, like the repeating charges along its backbone, are essential; so-called forced moves in the space of available design (see Dennett 1995, 128–131). However, with the case of the familiar

genetic alphabet, it seems that nature *could have* chosen otherwise, but for some reason it simply did not. It seems to be a partly contingent solution. Because it is not possible for evolution to change this situation anymore, the only plausible way to study these questions is to use synthetic design methods. This is also good way to naturalize the notion of biological possibility: To show that how-possibly models and speculative scenarios of evolutionary theory and the rest of biology can be given a philosophically satisfying reading.

References

Bayer, T.S. 2010. “Using Synthetic Biology to Understand the Evolution of Gene Expression.” *Current Biology* 20(17): R772–R779.

Bechtel, W., and R.C. Richardson. 1993. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton, NJ: Princeton University Press. 2nd ed. 2010. Cambridge, MA: The MIT Press.

Benner, S.A., Z. Yang, and F. Chen. 2011. “Synthetic Biology, Tinkering Biology, and Artificial Biology. What Are We Learning?” *Comptes Rendus Chimie* 14(4): 372–387.

Craver, C.F. 2006. “When Mechanistic Models Explain.” *Synthese* 153(3): 355–376.

———. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Craver, C.F., and L. Darden. 2013. *In Search of Mechanisms: Discoveries across the Life Sciences*. Chicago: The University of Chicago Press.

Dawkins, R. 1986. *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe without Design*. Norton & Company.

Dennett, D.C. 1995. *Darwin’s Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon and Schuster Paperbacks.

Elowitz, M.B., and W.A. Lim. 2010. “Build Life to Understand It.” *Nature* 468(7326): 889–890.

Endy, D. 2005. “Foundation for Engineering Biology.” *Nature* 438(7076): 449–453.

Knuuttila, T., and A. Loettgers. 2013. “Synthetic Modeling and Mechanistic Account: Material Recombination and Beyond.” *Philosophy of Science* 80(5): 874–885.

Kwok, R. 2012. “DNA’s New Alphabet.” *Nature* 491(7425): 516–518.

Malyshev, A., K. Dhami, T. Lavergne, T. Chen, N. Dai, J.M. Foster, I.R. Corrêa Jr, and F.L. Romesberg. 2014. “A Semi-Synthetic Organism with an Expanded Genetic Alphabet.” *Nature* 509(7500): 385–388.

Marlière, P., J. Patrouix, V. Döring, P. Herdewijn, S. Tricot, S. Cruveiller, M. Bouzon, and R. Mutzel. 2011. “Chemical Evolution of a Bacterial Genome.” *Angewandte Chemie International Edition* 50(31): 7109–7114.

Morange, M. 2009. “Synthetic Biology: A Bridge between Functional and Evolutionary Biology.” *Biological Theory* 4(4): 368–377.

Rosenberg, A. 2006. *Darwinian Reductionism: Or, How to Stop Worrying and Love Molecular Biology*. Chicago: The University of Chicago Press.

Schmidt, M. 2010. “Xenobiology: A New Form of Life as the Ultimate Biosafety Tool.” *BioEssays* 32(4): 322–331.

Schrödinger, E. 1944. *What Is Life?* Cambridge: Cambridge University Press.

Sprinzak, D., and M.B. Elowitz. 2005. “Reconstruction of Genetic Circuits.” *Nature* 438: 443–448.

Switzer, C., S.E. Moroney, and S.A. Benner. 1989. “Enzymatic Incorporation of a New Base Pair into DNA and RNA.” *Journal of the American Chemical Society* 111(21): 8322–8323.

Szathmáry, E. 2003. ”Why Are There Four Letters in the Genetic Alphabet?” *Nature Reviews Genetics* 4: 995–1001.

Thyer, R., and J. Ellefson. 2014. “New Letters for Life’s Alphabet.” *Nature* 509(7500): 291–292.

Wagner, A. 2005. *Robustness and Evolvability in Living Systems*. Princeton, NJ: Princeton University Press.

Wimsatt, W.C. 2007. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge, MA: Harvard University Press.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Connecting Inquiry and Values in Science Education: An Approach based on John Dewey's Perspective

Introduction

The results of science surround us and structure our everyday world, and science impacts us almost every moment in our lives. We make numerous decisions on issues related to science during our lifetime, and every time we make such a decision, values are involved, because science is value-laden (Anderson, 2004; Biddle, 2013; Brown, 2012; Douglas, 2000, 2009; Kourany, 2010; Longino, 2002). In fact, values play a role not only in decision-making of science-related issues such as socio-scientific issues, but also in science practice. As Coulo (2014) pointed out, science not only bears on our values in many ways, but science is also affected by values because ethical and political responsibilities of scientific work and knowledge impact scientists and science. Values play an implicit role in the choice of research subjects and research methods (Coulo, 2014). Also individual scientists may choose to engage in certain kinds of research, but different societies and institutions may encourage or discourage them (Forge, 2008). Furthermore, these non-epistemic types of values including ethical, social, and political responsibilities affect science practice because of inductive risk (Douglas, 2000). Therefore teaching and learning about the role of values in science in socio-scientific and controversial issues can play a role in humanizing sciences and illustrating their ethical, cultural and political facets (Matthews, 1994).

A study by Evagorou, Jimenez-Aleixandre, and Osborne (2012) showed an example of how non-epistemic values affected students' decision-making in socio-scientific issues and how little scientific inquiry affected decision-making. When two groups of students with different background were asked to make a decision on a socio-scientific issue, their decisions appeared to be based on their cultural and social background rather than the inquiry that they conducted in the science class. There was little change in their opinions before and after the class, and even though they conducted an inquiry based on various related information, students tended to accept only supporting evidence to their opinions. Students' reasoning for their decisions was not evidence-based (Evagorou et al., 2012). Another study by Nielson (2012) showed that students co-opted science to make it appear that their evaluative claims were solidly supported. Furthermore, students used scientific evidence not only for justifying their claim but also for emphasizing the importance of their claim (Nielsen, 2012).

These are a few of the examples showing that conducting scientific inquiry does not automatically help students make an informed decision using inquiry-based evidence. Scientific inquiry has been emphasized in science education because it is expected to help students understand, evaluate and make an informed decision for science-related issues (American Association for the Advancement of Science [AAAS], 1993; Rutherford & Ahlgren, 1990). K-12 science education has focused on educating all citizens, and people who are well educated in science, whether they are scientists or non-scientists, are expected to possess scientific habits of mind, be capable of engaging scientific inquiry, and to reason well in scientific contexts (National Research Council [NRC], 2012). Overall, they are expected to make an informed

decision when they face a controversial science-related issue. Doing scientific inquiry in the science class, however, seems not become a useful experience for students to make a decision in socio-scientific issues as expected.

In this paper, we explore how to help students use inquiry in decision-making based on John Dewey's perspective. Science education owes a lot to John Dewey's ideas of how science should be viewed and what science education should do (Wong et al., 2001). Unfortunately, although Dewey's ideas can be found in every facet of progressive science education in America and in the international science education, they have been underappreciated or misunderstood in many ways (Wong et al., 2001). Therefore, it is worth returning to Dewey's perspective of inquiry in science and exploring how it is related to decision-making.

A Missing Link in Science Education Standards

Inquiry is central to science learning and a prominent feature of science education standards including *National Science Education Standards* (NRC, 1996), *Inquiry and the National Science Education Standards* (NRC, 2000), and *Benchmarks for Science Literacy* (AAAS, 1993) focus on scientific inquiry. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts and Core Ideas* (NRC, 2012), and the *Next Generation Science Standards* (NGSS Lead States, 2013) also emphasize inquiry through "science and engineering practices" dimension. Meanwhile, decision-making is another important feature that has been emphasized in science education standards.

> In a world filled with the products of scientific inquiry, scientific literacy has become a necessity for everyone. Everyone needs to use scientific information to make choices that arise everyday (NRC, *National Science Education Standards*, 1996, 1p).

> We believe that the education of the children of this nation is a vital national concern. The understanding of, and interest in, science and engineering that its citizens bring to bear in their personal and civic decision making is critical to good decisions about the nation's future (NRC, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts and Core Ideas,* 2012, viii).

Making a good decision using scientific information in personal or civic issues is undoubtedly important so that students should learn it through science education. The term "inquiry" is used in two different ways in science education. First, it refers to the abilities and understanding students should develop to be able to conduct scientific investigations and second, it refers to the teaching and learning strategies (NRC, 2000). If inquiry also refers the teaching and learning strategies, it implies that inquiry can be used to learn an informed decision-making.

How conducting inquiry helps students learn an informed decision-making, however, is not explicitly explained in science education standards. Instead, *Benchmarks for Science Literacy* (AAAS, 1993) mentions critical response skills that students need to learn to make judgments based on what they know in science. According to this standard, how to use supporting evidence, the language, and the presented argument is an important skill to make judgments of whether taking the claim in question seriously or not, so students should learn such a skill and practice it to make it a lifelong habit of mind.

> Apart from what they know about the substance of an assertion, individuals who are science literate can make some judgments based on its character. The use or misuse of supporting evidence, the language used, and the logic of the argument presented are important considerations in judging how seriously to take some claim or proposition. These critical response skills can be learned and with practice can become a lifelong habit of mind (AAAS, 1993, 298p).

Learning critical response skills is not, however, enough for students to learn an informed decision-making. First, critical response skills are only for making judgments to accept some claims, and decision-making requires more than a judgment to accept the claim or the proposition. For example, in every decision-making, values are involved. Without considering involved values, accepting a certain claim does not automatically achieve a decision. Second, critical response skills mentioned in the standards are skills to judge a given claim or preposition, not skills to use or learn to do the inquiry. Therefore there is a missing link between scientific inquiry and decision-making. If we do not know how conducting scientific inquiry helps students make an informed decision in science-related issues, the first question we need to explore will be how scientific inquiry is related to decision-making. We explored this question based on Dewey's views of the relationship among scientific inquiry, value judgment in science, and decision-making.

Scientific Inquiry, Value Judgment, and Decision-Making

Scientific inquiry and its contribution to society play a central role in the philosophical and educational work of John Dewey. Dewey (1910/1995) emphasized that science is not only a subject-matter and body of results, but also a process or method. He pointed out that science

education focused too much on teaching a body of ready-made knowledge and not enough on inculcating a method of thinking, in other words, scientific inquiry (Dewey, 1910/1995). For Dewey, the primary goal of science education is to develop students' ability to inquire as a habit of mind. Dewey's emphasis on scientific inquiry is similar to the emphasis made in *Benchmarks for Science Literacy* (AAAS, 1993), *National Science Education Standards* (NRC, 1996), and *A Framework for K-12 Science Education: Practices, Crosscutting Concepts and Core Ideas* (NRC, 2012). Today, the major goal of science education is for students to achieve science literacy, and scientifically literate people are expected to be able to make informed decisions on the science-related issues that they face in their lives (AAAS, 1993; NRC, 1996, 2012; Rutherford & Ahlgren, 1990). Thus, whether experiences of conducting scientific inquiry can help students in making informed decisions should be the important question to explore. Decision-making requires value judgment. Then the relationship between scientific inquiry and value judgment needs to be examined to explain how scientific inquiry can help students make a decision. *Science for All Americans* described scientific inquiry, values and attitudes as habits of mind (Rutherford & Ahlgren, 1990). Although these concepts were considered as essential, they were only presented in a way that juxtaposed them as separate and independent factors. What seems to be missing here is the connection between scientific inquiry and value judgment. This is the place that John Dewey's idea of scientific inquiry and of the relationship between inquiry and values can be used to make the missing connection.

According to John Dewey, the uses of scientific inquiry can improve students' ability to make value judgment (Webster, 2008). Inquiry and values are not separate but related because

the direction taken by inquiry is under the influence of values (Dewey, 1948a; 1948b). Thus, in science, inquiry should not be guided by inappropriate, external interests as Dewey explained below (Dewey, 1948a).

> The actual course of scientific inquiry has shown that the best interests of human living in general, as well as those of scientific inquiry in particular, are best served by keeping such inquiry "pure" from interests that would bend the conduct of inquiry to serve concerns alien to conduct of knowing as its own end and proper terminus (Dewey, 1948a, p.206).

"Pure" inquiry does not mean value-free ideal in scientific inquiry. Rather, it means that, when scientific inquiry is not misguided by inappropriate interests, it works based on evidence-based thinking, critical thinking and open evaluation, and eventually, it can contribute to make judgments as intellectual as possible (Dewey, 1910/1995; Webster, 2008). The inappropriate, external interests, the "concerns alien to conduct of knowing as its own end and proper terminus," are not all non-epistemic values, but rather, those values arrived at prior to and dogmatically held independently of scientific inquiry. Dewey (1910/1995) warned that if science is succumbed to inappropriate, external interests, it is no longer able to contribute to social and moral ideals, and further, to democracy.

> The modern warship seems symbolic of the present position of science in life and education. The warship could not exist were it not for science: mathematics, mechanics, chemistry, electricity supply, the technique of its construction and management. But the aims, the ideals in whose service this marvelous technique is displayed are survivals of a pre-scientific age, that is, of barbarism. Science has as yet had next to nothing to do with forming the social and moral ideals for the sake of which she is used (Dewey, 1910/1995, p.397).

The military interests behind the warship are precisely the kind of inappropriate, dogmatic, prescientific values that Dewey hopes to keep out of science, in favor of values produced or tested in the course of scientific inquiry. In fact, "when the actual courses of scientific inquiry has shown the best interests of human living (Dewey, 1948a, p.206)," scientific inquiry can contribute to social and moral ideals (Dewey,1910/1995). Therefore Dewey argued that science should focus on what we should do, and not merely on how we would do it (Dewey, 1910/1995). Thinking about what we should do indicates value-laden thinking. So Dewey's argument implies that science is value-laden practice, so making "pure" scientific inquiry should include making a good value judgment. Figure 1 shows the relationship among scientific inquiry, value judgment, and decision-making based on Dewey's view. Values are involved in conducting scientific inquiry, so scientific inquiry needs to include making a good value judgment. In other words, conducting scientific inquiry is a value-laden activity, so making a good scientific inquiry can improve students' ability to make a good value judgment. Thus, Dewey's idea of scientific inquiry and of the relationship between inquiry and values contributes to make the missing connection between scientific inquiry and value judgment in science education standards. Based on Dewey's view, we can see now how scientific inquiry can contribute to make informed decisions. Decision-making requires value judgment. Scientific inquiry can improve the ability to make a value judgment. Therefore scientific inquiry can contribute to make an informed decision through value judgment.

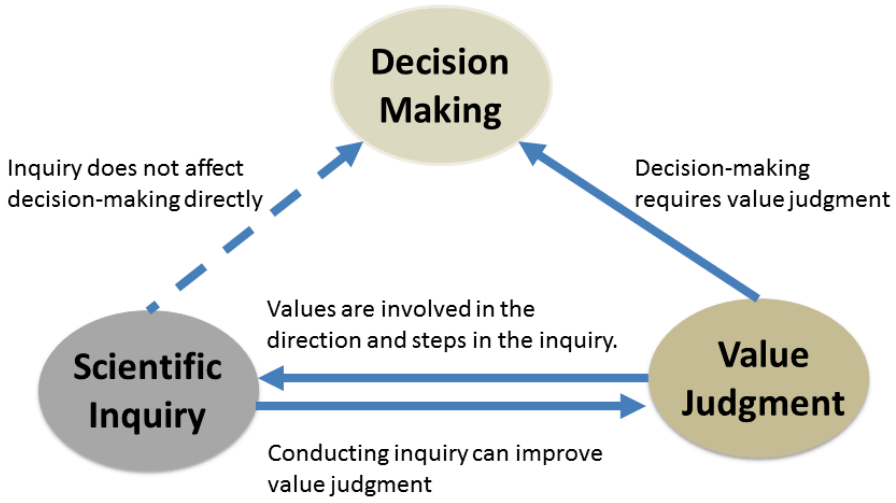


Figure 1. The relationship among scientific inquiry, value judgment, and decision-making in Dewey's view

Practical Value Judgment in Scientific Inquiry

The next question to explore will be how scientific inquiry can improve value judgment. Dewey argued that enforcing obedience to precepts does not do any good to students because it cut off the possibility of learning better ways to live by experimenting with them (Anderson, 2014). Considering Dewey's argument, it would not be appropriate to ask student to accept certain values as precepts when they conduct scientific inquiry, because it will take away the opportunity to do the experiment with various values. Students need to know that various values

can be involved during the inquiry and science education should provide an opportunity to students to conduct inquiries with those various involved values. Dewey suggested that a judgment of value is actually a case of a practical judgment, a judgment about the doing of something.

> A practical judgment has been defined as a judgment of what to do, or what is to be done: a judgment respecting the future termination of an incomplete and in so far indeterminate situation. To say that judgment of value fall within this field is to say two things: one, that the judgment of value is never complete in itself, but always in behalf of determining what is to be done; the other, that judgments of values (as distinct from the direct experience of something as good) imply that value is not anything previously given, but is something to be given by future action, itself conditioned upon (varying with) the judgment (Dewey, 1916, p.230).

The value judgment that students make during the scientific inquiry is also a practical judgment, because, at each step of the inquiry, students need to decide what to do or what is to be done, and values related in that situation will influence the decision. According to Dewey (1916), value judgment can be empirically tested (Anderson, 2014). When students make a value judgment to guide an action, there will be consequences of that particular action, and these consequences will determine if a certain judgement of values is appropriate or not. If students are aware that values are demonstrated in the judgment to guide an action, they can evaluate the values involved in the judgment by evaluating the consequences of the action. Thus, students' value judgment can be empirically tested while they are conducting the scientific inquiry. The uses of scientific inquiry can improve students' ability to make a value judgment (Webster, 2008).

Making a good value judgment can also help scientific inquiry. As a practical judgment, value judgment will be made during the whole process of scientific inquiry. Every time a student decides what to do, values will be involved in that decision of action, whether it is about selecting a particular method, collecting data or interpreting the results. Often, non-epistemic values such as ethical, social, and cultural values are considered to only affect external part of science practice, for example, the selection of hypotheses, restrictions on methodologies, and the use of scientific technologies (Douglas, 2000). These values, however, can also affect internal part of science practice such as statistical significance, evidence characterization, and interpretation of the results, because of inductive risk (Douglas, 2000). This is why science education includes value judgment in scientific inquiry because values affect both external part and internal part of the inquiry that students conduct. For example, social, ethical, or cultural values can influence the selection of hypotheses, so taking these values into account when selecting hypotheses can help students balance open-mindedness with skepticism (AAAS, 1993). Values can also influence in making a methodological choice. Exploring involved values and making value judgments can reduce the chances of choosing methodological options which have ethically unacceptable consequences (Douglas, 2000). Value judgment can also help in evidence characterization, when deciding how to characterize ambiguous data. Questioning and challenging values which might be involved in evidence characterization may help reduce possible errors in dealing with ambiguous data (Douglas, 2000). Value judgment can also help in the interpretation of the results. Not only epistemic values but also non-epistemic values may influence when interpreting the results. Taking a process to evaluate values when interpreting the

results of inquiry will be useful in avoiding interpretational mistakes (Douglas, 2000). Figure 2 shows how value judgment is involved during the scientific inquiry. Values influence the external part of the scientific inquiry such as the direction taken by the inquiry (Dewey, 1948a; 1948b), and are involved in the internal part of the scientific inquiry through practical judgments (Dewey, 2016; Douglas, 2000). Finally values are demonstrated in the judgment made during the inquiry (Brown, 2012; Webster, 2008).

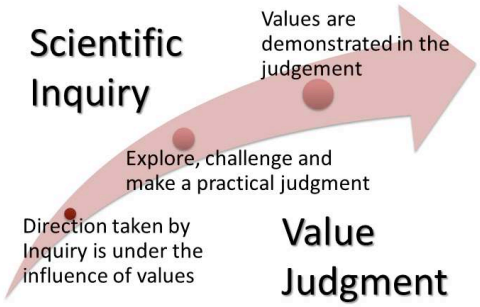


Figure 2. Making value judgment in scientific inquiry

Returning to John Dewey's Perspective

In Dewey's perspective, scientific inquiry and value judgment are closely related to each other. Relating inquiry and values, however, is not unfamiliar idea in science education, because *Science for All Americans* (Rutherford & Ahlgren, 1990) already recognized the interaction between values and science.

> Throughout history, people have concerned themselves with the transmission of shared values, attitudes, and skills from one generation to the next. Even today, it is evident that family, religion, peers, books, news and entertainment media, and general life experiences

> are the chief influences in shaping people's views of knowledge, learning, and other aspects of life. Science, mathematics, and technology can also play a key role in the process, for they are built upon a distinctive set of values, they reflect and respond to the values of society generally, and they are increasingly influential in shaping shared cultural values. Thus, to the degree that schooling concerns itself with values and attitudes, it must take scientific values and attitudes into account when preparing young people for life beyond school (Rutherford & Ahlgren, 1990, p.171).

This recognition, however, faded away in *Benchmarks of Science Literacy* (AAAS, 1993) the following publication after *Science for All Americans* (Rutherford & Ahlgren, 1990). *Benchmarks of Science Literacy* (AAAS, 1993) suggested practical standards for different age groups under the concepts and ideas from *Science for All Americans* (Rutherford & Ahlgren, 1990). There, honesty, curiosity, and balancing open-mindness with skepticism were suggested as scientific values that students should know.

> Honesty is a desirable habit of mind not unique to people who practice science, mathematics, and technology… Curiosity does not have to be taught. The problem is the reverse: how to avoid squelching curiosity while helping students focus it productively… [and] Balancing open-mindness with skepticism may be difficult for students (AAAS, 1993, p.284).

These are descriptions of epistemic values or epistemic virtues shared in science domain, not explanations of how values and science are related. Thus the relationship between values and science was introduced once, but was not pursued further, particularly not to the point of teaching value judgment as part of inquiry. Instead, students were asked to accept values like honesty, curiosity, and balancing open-mindness with skepticism as a sort of precepts. As Dewey pointed out, giving precepts without opportunities to examine them does not do any good to students in science education (Anderson, 2014). Instead of introducing "scientific values" as

precepts, scientific inquiry should provide both intellectual and methodological means to critically evaluate various values based on the idea in *Science for All American* (Rutherford & Ahlgren, 1990) and the idea of John Dewey (Dewey, 2016; Anderson, 2014).

Returning to Dewey's view of inquiry and values can help connecting a missing link between inquiry and values in science education. Table 1 shows a few problems that we recognized in current K-12 science education through *Benchmarks for Science Literacy* (AAAS, 1993), *National Science Education Standards* (NRC, 1996), and *A Framework for K-12 Science Education: Practices, Crosscutting Concepts and Core Ideas* (NRC, 2012). The missing link is that there is not an explicit explanation about how conducting scientific inquiry can help make informed decisions in science-related issues. One way to solve this problem is to explore the relationship between inquiry and values in science because decision-making requires value judgment. The connection between inquiry and values in science, however, are not explicitly explained either. Values in science are not supposed to be explored during the scientific inquiry, and that does not help connecting inquiry and values in science education. Table 1 also shows possible solutions to these problems, based on Dewey's view. According to Dewey (1916; 1948a), scientific inquiry should include good value judgments, and a value judgment in the scientific inquiry is a practical judgment to guide an action which result reflects involved values. Thus, conducting the scientific inquiry can improve students' ability to make a value judgment (Webster, 2008). Considering these ideas, students should be aware of a few things when they conduct inquiry in the science classroom. First, students should know that various values are involved in the scientific inquiry, and those values can be challenged and evaluated. Second,

students should know that they are making a practical value judgment at every step of the scientific inquiry, and they can evaluate the involved values by examining the result of the action. Third, students should know that conducting science inquiry needs to include a good value judgment. Then, connecting inquiry and values in science education can be completed, and the missing link among scientific inquiry and informed decision-making will eventually be connected in science education.

Problems in current science education	**Solutions based on Dewey's view**
Making inquiry does not automatically help making an informed decision.	Decision-making requires value judgment, and making inquiry can improve value judgment. If conducting scientific inquiry includes making a good value judgment, it can eventually help an informed decision-making.
Inquiry and values in science are not explicitly connected.	The direction of the inquiry is under the influence of values. During the inquiry, value judgment has to be a practical judgment, a judgment guiding an action. So the result of the inquiry will include the result of value judgment, and demonstrate involved values.
Values are provided as precepts and not explored during the inquiry.	Making a practical judgment during the scientific inquiry gives students an opportunity to critically evaluate various values and apply them. At each step of the inquiry, students will decide what to do after evaluating various values involved.

Table 1. Problems found in science education and solutions based on Dewey's view

Conclusion

Although there have been more than nine definitions about science literacy through history of science education (DeBoer, 2000), science literacy has been considered as an important goal for students to achieve (AAAS, 1993; NRC, 1996; 2012, Rutherford & Ahlgren, 1990). There is

a certain consensus of describing scientifically literate people as being familiar with the natural world, understanding some of the key concepts and principles in science, having a capacity for scientific ways of thinking, and being able to use scientific knowledge and ways of thinking for personal and social purposes (DeBoer, 2000). Also, scientific inquiry always has been one of essential attributes to achieve science literacy. Naturally, a scientifically literate person is expected to be able to make informed decisions for science-related issues based on inquiry. What is missing there, however, is that it has not been clear how scientific inquiry can contribute to make informed decisions. Since making decisions requires value judgment, the problem turned into what the relationship is between scientific inquiry and value judgment.

John Dewey's view that the uses of scientific inquiry can improve students' ability for value judgment provides that missing link between inquiry and decision-making. Inquiry is an active process of knowing by understanding, evaluating, and forming the knowledge. Learning science through inquiry transforms our world view by opening up for action (Kruckeberg, 2006). Inquiry also includes value judgment that is a practical judgment of guiding an action. Therefore, each step of scientific inquiry involves value judgment to decide what to do, and in this way, values influence both external and internal part of science practice. With the help of John Dewey's view, scientific inquiry in K-12 science education can be connected to value judgment, and eventually to decision-making. There, students can learn how to conduct the scientific inquiry and how to make practical judgment during the scientific inquiry. Reflecting the result of their practical judgment, students can evaluate values involved in their decision during the inquiry. Students can learn value judgment while conducting the scientific inquiry and these learning

experiences will help them when they make a personal or civic decision in science-related issues.

Ultimately these learning experiences will lead students to achieve science literacy.

References

Anderson, E. (2004). "Uses of Value Judgments in Science: A General Argument, With Lessons from a Case Study of Feminist Research on Divorce." Hypatia, 19(1): 1–24.

Anderson, E. (2014). "Dewey's Moral Philosophy", The Stanford Encyclopedia of Philosophy (Spring 2014 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu/archives/spr2014/entries/dewey-moral/

American Association for the Advancement of Science. (1993). Benchmarks for science literacy. New York: Oxford University Press.

Biddle, Justin (2013). "State of the Field: Transient Underdetermination and Values in Science." Studies in History and Philosophy of Science, 44(1): 124–133.

Brown, M. J. (2012). John Dewey's logic of science. HOPOS: The Journal of the International Society for the History of Philosophy of Science, 2, 258-306.

Coulo, A. C. (2014). Philosophical dimensions of social and ethical issues in school science education: values in Science Classrooms. In M.R. Matthews(ed.), International Handbook of Research in History, Philosophy and Science Teaching (pp 1087-1117). Dordrecht: Springer.

DeBoer, G. E. (2000). "Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. Journal of Research in Science Teaching, 37, 582-601.

Dewey, J. (1948a). Common sense and science: Their respective frames of reference. The Journal of Philosophy, 45, 197-208.

Dewey, J. (1948b). Reconstruction in philosophy. Boston: Beacon Press.

Dewey, J. (1910/1995). Science as subject-matter and method. Science & Education, 4, 391-398. (Original work published 1910)

Douglas, H. (2000). Inductive risk and values in science. Philosophy of Science, 67, 559-579.

Douglas, H. (2009). Science, Policy, and the Value-Free Ideal. Pittsburgh: University of Pittsburgh Press.

Evagorou, M., Jimenez-Aleixandre, M. P. & Osborne, J. (2012). Should we kill the grey squirrels? A study exploring students' justifications and decision-making. International Journal of Science Education, 34, 401-428.

Forge, J. (2008). The responsible scientist. Pittsburgh, PA: University of Pittsburgh Press.

Kourany, J. A. (2010). Philosophy of Science after Feminism. Oxford University Press.

Kruckeberg, R. (2006). A Deweyan perspective on science education: Contstructivism, experience, and why we learn science. Science & Education, 15, 1-30.

Longino, Helen E (2002). The Fate of Knowledge. Princeton University Press.

National Research Council (1996). National science education standards. Washington DC: National Academy Press.

National Research Council (2000) Inquiry and the national science education standards. Washington DC: National Academy Press.

National Research Council (2012). A framework for k-12 science education: Practices, crosscutting concepts, and core ideas. Washington DC: National Academy Press.

Nielsen, J. A. (2012). Co-opting Science: A preliminary study of how students invoke science in value-laden discussions. International Journal of Science Education 34, 275–299.

NGSS Lead States (2013). Next generation science standards: For states, by states. http://www.nextgenscience.org/

Rutherford, F. J. & Ahlgren, A. (1990). Science for All Americans. New York: Oxford University Press.

Webster, S. (2008). How a Deweyan science education further enables ethics education. Science & Education, 17, 903-919.

Wong, D., Pugh, K. & the Dewey Ideas Group at Michigan State University. (2001). Learning science: A Deweyan perspective. Journal of Research in Science Teaching, 38, 317-336.

Draft Symposium Paper for PSA 2016
Please do not quote without permission

Towards Mechanism 2.0: Expanding the Scope of Mechanistic Explanation

Arnon Levy
Hebrew University of Jerusalem

William Bechtel
University of California, San Diego

1. Introduction

Accounts of mechanistic explanation, especially as applied to biology and sometimes going under the heading of "new mechanism," provided an attractive alternative to nomological accounts that preceded them. These accounts were motivated by selected examples, drawn primarily from cell and molecular biology and neuroscience,. These examples pointed to sharp contrasts between real-life biological explanation and discovery and the then-dominant models of scientific explanation. However, the range of examples that scientists take to be mechanistic explanations is far broader. We focus on examples that differ from those traditionally recruited by Mechanists. Our contention is that attention to additional examples will lead to a richer conception of mechanistic explanation, prompting a shift from what we refer to as Mechanism 1.0 to Mechanism 2.0. In suggesting such a move, our goal is not to downplay the importance of Mechanism 1.0 and the progress it signified. Mechanism was a big step forward in philosophy of biology. We just think it's time to take the next step. Furthermore, by adopting the language of Mechanism 1.0 and 2.0 we mean to signal that we anticipate further enhancements to the conception of mechanistic explanation as philosophers of science attend to more examples of scientists advancing what they characterize as mechanistic explanations.

One way to approach the distinction between Mechanism 1.0 and 2.0 is to return to the machine metaphor that inspired mechanistic research in biology and was invoked explicitly by Bechtel and Richardson (1993/2010) and implicitly - primarily in the choice of examples - by other writers on mechanistic explanation. Most mechanists have attempted to differentiate biological mechanisms from machines. However, the examples used to motivate accounts of mechanistic explanation by, for example, Bechtel & Richardson (1993/2010), Bechtel (2006), Machamer, Darden, and Craver (2000), and Craver and Darden (2013) are in fact much like traditional machines. Thus, protein synthesis is presented as involving the creation of mRNA from the DNA template in the nucleus and the transport of the mRNA to the ribosome, where it serves as a template for forming a chain of amino acids. Oxidative metabolism is described as localized to the mitochondria of cells where a specific set of enzymes catalyze the successive oxidation of metabolites until only carbon dioxide and water remain, generating ATP in the process. In these examples, the

mechanism consists of a bounded set of enduring entities or parts in a fixed configuration. These explanations accord central importance to the structure of parts and often envision the mechanism's organization as sequential, or perhaps branching. As in classical machines—steam engines, typewriters and food processors—the parts are envisaged as performing the same activities or operations every time they are called upon so as to produce the phenomenon to be explained.

These features of the examples, we contend, did much to advance an attractive and compelling picture of explanation that attracted much interest. For example, they established that such explanations did not fit the D-N model. By portraying a sequence of operations that resulted in storing energy in ATP or synthesizing proteins, scientists explained these processes without explicitly invoking laws. Moreover, for philosophers who were expanding their focus beyond justification to include discovery, these examples provided case studies of how the two practices were connected (Darden, 2006). The account also offers norms of success: to understand how a system in nature works, one should be able to identify its parts, demonstrate what operations they perform, and describe how they are organized so as to work together. If researchers cannot identify parts and trace how they operate on each other, they fail to show how the mechanism generated the phenomenon in terms of parts.

Still working within the framework of Mechanism 1.0, some philosophers began to focus on biological mechanisms whose parts are organized in a more complex manner (e.g., into multiple feedback loops). This undercut the ability of scientists to mentally rehearse the functioning of the mechanism to understand how it brought about its behavior. Instead, scientists had to appeal to mathematical representations and perform computational simulations. When the required mathematical representations are non-linear, computational simulations show how mechanisms can produce complex behavior (e.g., oscillations that partly synchronize with each other). While mechanists such as Bechtel and Abrahamsen (2010) and Brigandt (2013) distinguished such explanations as dynamical mechanistic explanations, they did not fundamentally challenge the framework of mechanism 1.0: they still appealed to a stable and bounded set of parts whose structure determines their interactions. Rather, they took a relatively small step away from Mechanism 1.0 to version 1.1.

Our project is much the same as the philosophers who advanced Mechanism 1.0. We focus on examples of explanation in biology. The difference is that we focus on ones that do not fit the picture of Mechanism 1.0 or 1.1. Some philosophers might see these departures from Mechanism 1.0 as requiring abandoning mechanism altogether. We certainly think that there are explanations that are not mechanistic – such as teleological, etiological and perhaps mathematical explanations. But the cases we explore here, while differing from Mechanism 1.0 in specific respects that we will highlight, are still recognizably parts-and-operations explanations, and are typically characterized as such by scientists. We will identify several ways in which these examples reveal limits of Mechanism 1.0 in the next section. In our view, these examples motivate expanding and reconceptualizing what a

mechanism is (hence, we speak of Mechanism 2.0). Although we are not yet at the point where it makes sense to offer a full characterization of Mechanism 2.0, in section 3 we will both articulate how the examples offered for Mechanism 2.0 enrich our understanding of mechanistic explanation and how mechanisms may differ from machines as they have been traditionally understood. We will also identify work that remains to be done in developing the conception of Mechanism 2.0.

2. The limits of Mechanism 1.0

In this section we will discuss departures from five key aspects of Mechanism 1.0: the appeal to the structural features of parts; the idea of a stable and straightforward organization; the assumption that parts are stable; the idea that mechanisms have well-defined boundaries in space and time; and the conception of mechanisms themselves as enduring entities.

2.1. Parts that are not discrete entities

In the examples used to illustrate and motivate Mechanism 1.0, mechanisms consisted of discrete entities that could be identified structurally in terms of properties such as shape, size, and mass. Indeed, many mechanistic explanations are presented in terms of individual entities such as molecules that undergo transformations (perhaps binding to another molecule and changing their shape in the process). (This is how mechanisms are represented in many diagrams of mechanisms, including those below.) But in fact the working part is very often not a single entity, but a large collection of similarly structured entities. And it is very often not only—sometimes not primarily—structural properties that matters but aggregative features such as the concentrations of these entities that performs the work.

Phenomena that involve electrical potentials over membranes provide examples whose explanations depend on concentrations. For instance, ATP synthesis strongly depends on the direction of the proton gradient across the mitochondrial membrane. Similarly for action potentials: to understand how an action potential works, you can't focus merely on the molecules that are involved (sodium, potassium, ion channels etc.) and their structures. It is the relative concentrations of ions, inside and outside the cell, that determine the timing and size of a spike. In such examples, components with the relevant structures are present throughout the process, on both sides of the membrane. What drives the process isn't their mere presence or structure. It is their relative concentration, which changes continuously as the mechanism operates. As protons build up in the mitochondria's inter-membranal space, they generate a potential, then is then converted into ATP via the ATP synthase "windmill." Structural aspects matter here, of course, but without careful attention to concentrations, the system cannot be understood. The role of concentration is perhaps even more subtle in action potentials. Sodium and potassium build up to a steady state concentration, maintaining a resting potential. When an above-threshold excitation arrives, sodium and potassium change concentrations quickly and in opposite "directions." It is the precise shifts in concentration and their timing that determines whether a spike is

generated. The structures of ions and ions channels, matter, but changes in concentrations are key.[1]

A major focus of contemporary biology is the regulation of biological processes, and here one often finds references to switches. A recent paper by Nathan (2014) looks at such a case in depth, arguing for a notion of causation by concentration. Nathan's analysis is compelling, though he does not explore potential ramifications for Mechanism. We focus on this aspect. A genetic switch is a bi-stable system in which a given gene can be either "on", leading to high levels of transcription, or "off", leading to low levels of transcription. The lac operon and the viral lambda switch are very well studied examples. In genetic switches, it is not merely the ability of an inducer or an inhibiting molecule to bind to DNA, and initiate transcription, that explains switching. Indeed, inducers and inhibitors are typically bound to DNA, to some extent, at all times. But what determines whether a switching event occurs is the relative rate of binding, which is determined by the their relative concentrations. Again, the structure of the parts remains constant. It is the shift in concentrations that moves the system into wholly new states. (For an overview of the principles underlying genetic switching, see Nelson, 2015, Chapter 10.)

The cases presented in this section illustrate that in some explanations what is doing the explanatory work is not the structure of the parts in question, but the concentrations of the parts. Ironically, both ATP synthesis and the generation of action potentials have been used in support of Mechanism 1.0. The fact that in these examples it was the concentration of protons or electrons that mattered was not noted. But it has direct consequences in terms of how the explanation in question is to be confirmed, what kind of discovery strategies it will be linked to and so on. Thus, a move away from the idea of parts as (merely) structural units has important consequences.

2.2. Organization that is not fixed

Although it was typically not commented on, those offering examples of Mechanism 1.0 recognized that, like parts of machines, the parts of a mechanism go through a sequence of different states as the mechanism generates a phenomenon. In many cases, including enzyme-catalyzed metabolic reactions, cell-to-cell signaling, DNA and RNA processing, and various other phenomena, the ability of a part to perform an operation depends on its three-dimensional shape and associated physical features such as its electrical charge. It is as a result of such features that an enzyme is able to bind to a substrate. When bound to a substrate, it is no longer able to bind to another and the conformation of the molecule is changed. Once the reaction is complete, it returns to its initial conformation and is able to bind another molecule of substrate.

---

[1] This is illustrated by the fact that Hodgkin and Huxley, who knew little about the structures of underlying molecules, still managed to produce a seminal advance in our understanding of action potentials via a model that took into account facts about concentrations alone. See Levy (2013).

[2] In the spirit of our discussion of concentrations above, registration of time of day appears to be a population

While recognizing these local changes of parts and how they interact with each other in the course of a mechanism's operations, Mechanism 1.0 presented the overall organization of mechanisms as unchanging. This is reflected in what is perhaps the standard form in which many mechanisms are represented: a flow diagram in which nodes represent parts and arrows represent operations performed by one part on other parts. Most diagrams do not represent how individual parts change structurally over time. For example, Figure 1, showing the activities of transcription and translation leading to the synthesis of new proteins, only shows tRNAs binding to amino acids and transporting them to the ribosome, not any change in the tRNA that results. Rather, the focus is on tracing how each part interacts with other parts. The various activities that are shown--the assembly of the mRNA along the DNA template, the transport of the mRNA (and tRNA and rRNA) to the cytoplasm, the ferrying of amino acids to mRNAs by tRNAs, and the binding of amino acids are into a polypeptide chain--are presented as enduring.

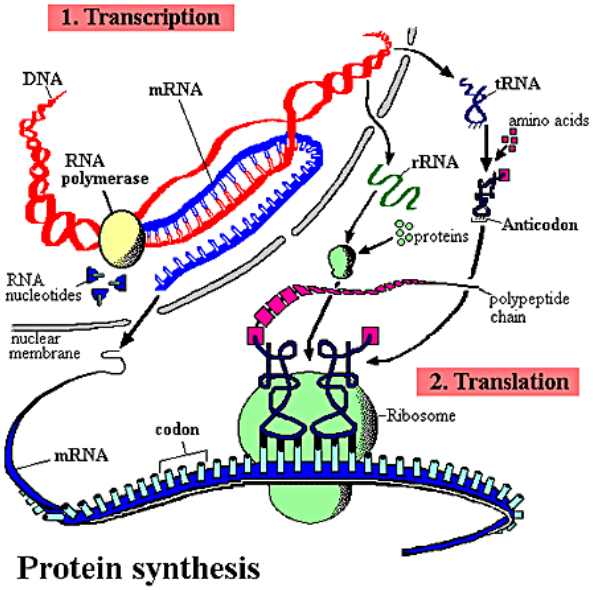


Figure 1.

In the case of some mechanisms, however, not only do the parts undergo changes but *which parts interact with which other parts changes* as the mechanism functions. The cyanobacterial circadian clock provides a relatively simple example. The core mechanism involves just three proteins, KaiA, KaiB, and KaiC and a source of ATP. The ATP provides phosphates that reversibly phosphorylate KaiC at two binding sites. Since one site is both phosphorylated and dephosphorylated first, the relative concentrations of KaiC in the different phosphorylation states uniquely specifies the time of day (Rust, Markson, Lane et al., 2007). (Note that in this case as well it is concentrations that do the work in the mechanism.) KaiC is itself capable of both autophosphorylation and

autodephosphorylation, and which operation it performs depends on KaiA and KaiB. KaiA can bind to KaiC in two different regions. When it acts alone, it binds to the A-loops coming out of the C2 domain of KaiC and fosters phosphorylation by changing KaiC's conformation (Figure 2, left). But when KaiB binds to the C1 domain, KaiA moves to bind to KaiB (Figure 2, right). By changing KaiC's conformation in a different manner, KaiA and KaiB promote dephosphorylation of KaiC (Tseng, Chang, Bravo et al., 2014). Even in this simple mechanism, KaiA interacts with different entities at different times, altering how the mechanism behaves. Which organization is implemented at a given time determines how still other parts (KaiC) behave and hence what time the clock registers. To exhibit this change in organization in diagram form, researchers often use separate diagrams to show the organization at different times.

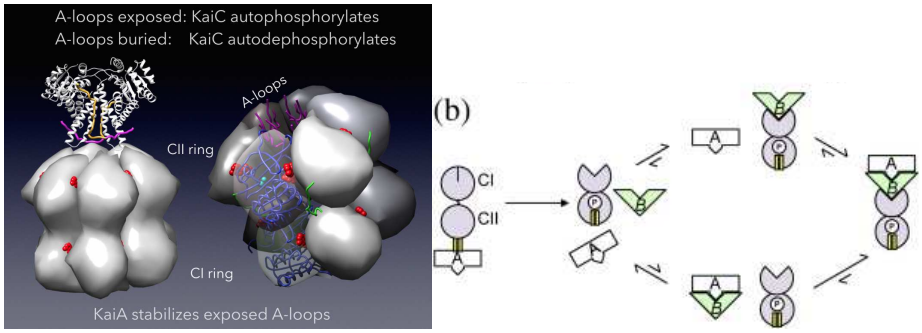


Figure 2. Left from Andy LiWang (http://faculty.ucmerced.edu/aliwang/galleries/image-gallery). Right: From Tseng, et al. (2014).

Examples such as this make clear that organization is a dynamic property of some mechanisms. The arrangement of parts, who interacts with whom when, varies as the mechanism functions, in part due to the operations the parts themselves are performing. So here we have another departure from Mechanism 1.0, inasmuch as it assumed a stable organizational pattern for mechanisms.

2.3. Mechanisms whose parts change over time

The examples we looked at so far involved changing concentrations and organizational patterns. But the "list" of parts playing a role in the mechanism was presumed stable. This is what we expect, based on the analogy to machines. But in some biological mechanisms, the parts change over time. Researchers have often failed to notice this since until recently it was not common to investigate a mechanism at different times. But as a consequence of automated data collection techniques, it has become possible to collect data about parts and their interactions at different times. For instance, in a study combining data about which proteins that can interact with each other to form complexes with time-series data on gene expression in yeast, de Lichtenberg, Jensen, Brunak et al. (2005) were able to provide evidence of how different parts are incorporated into a mechanism during different stages of the cell cycle. Although many genes are constitutively expressed, they

identified 600 genes (out of the approximately 6000 genes in yeast) that are only expressed during one stage of the cell cycle. Figure 3 shows one mechanism they investigated, the prereplication complex, which had previously been shown to involve Cdc28p and several Clb-type cyclins that function in regulating stages of the cell cycle. de Lichtenberg et al. demonstrated that individual cyclins are expressed and become available to bind with Cdc28p at different phases of the cycle. The color in Figure 3 shows the phase of the cycle in which the cyclins are synthesized: those shown in purple are expressed at the beginning of the M (mitosis) stage, those in orange through yellow during the G1 (gap 1) stage, those in green during the S (synthesis) stage, and those in blue during the G2 (gap 2) phase. At the end of the G2 phase the action of Cdh1p leads to the ubiquitination and degradation of the cyclins via Clb2p. The discovery that different parts are added to the mechanism at different stages of the cell cycle explains the different regulatory roles the mechanism plays at different stages.

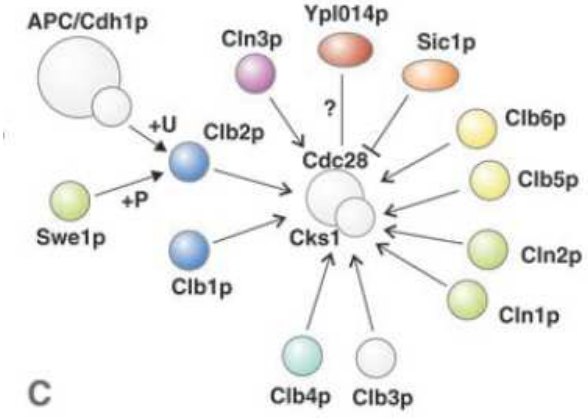


Figure 3. From de Lichtenberg et al. (2005).

The fact that the proteins that constitute the parts of some mechanisms change over time is not surprising. Biological mechanisms continually degrade and so are continually being built and repaired through the synthesis of new proteins. By not attending to when different parts are synthesized, Mechanism 1.0 tended to view mechanisms as enduring entities. Emphasizing their construction and degradation is thus a step beyond Mechanism 1.0.

2.4. Mechanisms with porous boundaries

When one buys a machine such as a toaster, it typically comes packaged in a box. The toaster cannot toast bread without electricity and bread being supplied, but the boundaries between the mechanism and the environment remain well delineated. Moreover, it operates in sharply distinguished time periods. Mechanism 1.0 made similar assumptions about biological mechanisms. An important discovery strategy was to try to localize the mechanism responsible for a phenomenon in time and space. These efforts often appeared

to be successful. Mammalian circadian researchers, for instance, introduced the term *clock* to designate the responsible mechanism and localized it initially in the suprachiasmatic nucleus of the SCN (Moore & Eichler, 1972) and subsequently in individual cells in the SCN (Welsh, Logothetis, Meister et al., 1995).[2] Treating circadian rhythms as produced within individual cells, investigators sought the component genes and proteins that constituted a feedback loops that created oscillations--a proteins accumulated, they fed back to inhibit their own expression (Reppert & Weaver, 2001) (Reppert and Weaver, 2001).

Researchers of course expected different mechanisms to have outputs that were used in other mechanisms--protein synthesis generated proteins that provide parts of other mechanisms. But over time researchers have found more and more interactions between mechanisms, so many that the notion of a mechanism as a well-delineated "thing", with specified boundaries in time and space, can come to look questionable.

We start with a specific example of an unexpected connection between mechanisms. As researchers investigated how circadian proteins fed back to alter the expression of their own genes, they discovered that one critical protein, CLOCK, affects gene expression as a histone acetyltransferase. In searching for a histone deacetylase needed to counterbalance CLOCK, researchers identified SIRT1, a molecule already known to be critical for a host of cellular activities including basic metabolism (Sahar & Sassone-Corsi, 2009; Bass & Takahashi, 2010) (left side, Figure 5). Parts identified as belonging to the mechanisms responsible for these activities also affect circadian rhythms.

As a result of sharing components, mechanisms affect the functioning of other mechanisms not just through their inputs and outputs, as characteristic of the examples advanced by Mechanism 1.0, but through many of their internal operations. The example of SIRT1 is just one of a host of discoveries researchers have made where parts of the clock mechanism interact with parts of other mechanisms involved in other cell functions. Using sRNA screens to identify genes that when modified affected clock performance, Zhang, Liu, Hirota et al. (2009) found many such genes involved in a wide array of other cell functions. On the right in Figure 5 core clock genes are shown in dark and light blue; those shown as connected in various ways to the core clock genes are the ones that altered the period of the clock when mutated and whose proteins are known to interact with core clock proteins. These are normally identified as parts of different cellular mechanisms.

[2] In the spirit of our discussion of concentrations above, registration of time of day appears to be a population level effect, but with the extra complication that there are mechanisms to promote synchrony in local populations and complex dynamics over the whole (Welsh, Takahashi, & Kay, 2010).

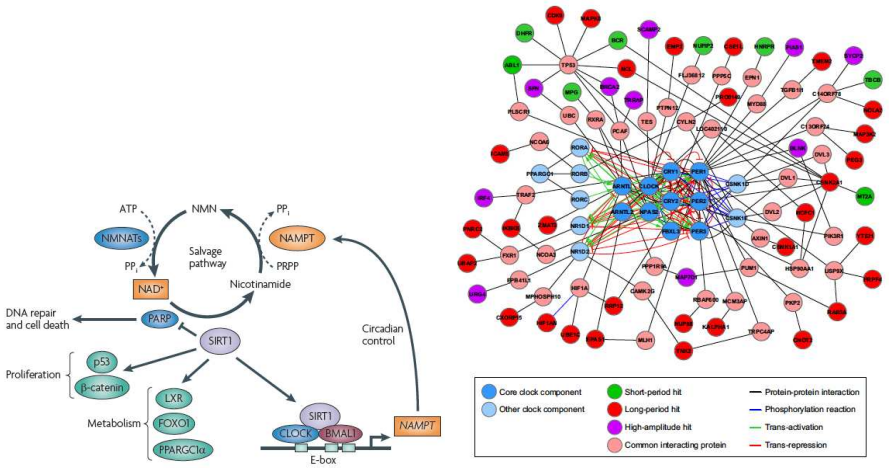


Figure 5. Left: Bass and Takahashi (2010). Right: Zhang et al. (2009)

These developments in circadian research exemplify a frequent trajectory of mechanistic research: in the wake of identifying a few parts of a mechanism, researchers continue to discover more and more entities that affect the functioning of the mechanism but are also recognized parts of other mechanisms. Recent efforts to represent the components of cells in networks (e.g., of protein interactions) often reveal that classically characterized mechanisms correspond to what are termed *modules*--clusters of parts that are more interconnected with each other than with other components, but which still have extensive connections to other modules. There typically are not clear boundaries around modules. Rather than finding sharp boundaries, researchers rather make pragmatic judgments as to where to draw boundaries (Bechtel, 2015).

The lack of sharp boundaries also affects the time window in which a mechanism carries out an operation. In standard portrayals of an action potential, the cell is at its resting potential until a stimulus arrives which depolarizes the cell. If depolarization exceeds a threshold, an action potential is generated. A recovery period follows in which the cell is first depolarized beyond the resting potential until it gradually returns to the resting potential where it resides until the next stimulus arrives. But in fact neurons fluctuate around the resting potential. Moreover, the effects of a single action potential can be demonstrated to affect these fluctuations, and hence the propensity to generate further action potentials up to minutes later, well after the action potential to reach its termination conditions (and may well have produced multiple additional action potentials). Marom (2010) argues that this shows that there is no characteristic timescale for action potentials. Nonetheless, researchers assume a timeframe for action potentials, ignoring these long transient effects.

These examples show that rather than well-delineated spatial and temporal boundaries between what is in and what is outside a mechanism, as are found with machines and is

suggested by the examples advanced by Mechanism 1.0, biological mechanisms often have porous boundaries. As a result of treating these boundaries as more fixed than they are, researchers are only able to account for phenomena approximately. When these shortcomings become important, researchers expand boundaries to include other parts and operations, but without abandoning the search for mechanistic explanation one cannot include everything. Selecting boundaries between multiple plausible candidates is an important challenge in Mechanism 2.0.

2.5. Mechanisms that exist only transiently

In Mechanism 1.0 mechanisms are viewed as present ready to operate when appropriate conditions arise. This corresponds to how we typically view machines, although when we adopt a long enough time horizon we recognize that machines are built, maintained for a period, and eventually decay, at which point they may be recycled. This occurs in biology far more frequently than suggested by Mechanism 1.0. Some biological mechanisms appear only to come into existence under specific conditions and are degraded when no longer needed.

Ideker and Krogan (2012) characterized *differential network biology* as a strategy in which network representations of gene or protein interactions identified under different conditions are contrasted to identify modules in yeast cells that only appear transiently. Employing annotations such as those provided by Gene Ontology, these modules can often be linked to mechanisms as identified in traditional molecular biology. The result is the identification of mechanisms that arise only in some conditions, presumably ones in which the phenomenon for which they are responsible is required by the cell.

Employing a version of this strategy, Bandyopadhyay, Mehta, Kuo et al. (2010) compared gene interactions in yeast growing under unperturbed conditions with those in which methyl methanesulfonate (MMS), a DNA-alkylating agent, was added to the medium. For each condition they created an epistatic microarray profile (E-MAP) that identified pairs out of a set of 418 selected genes that interact when mutated (that is, together they have effects on colony growth different from the product of their individual effects, as would be expected if they did not interact). Interactions are viewed as indicators that the proteins coded by the genes operate together in a mechanism. Bandyopadhyah et al. identified 1905 interactions in the untreated and 2297 in the MMS condition. Most of the interactions were only found under one of these conditions. They then created a differential E-MAP by subtracting the E-MAP for one condition from that for the other. This revealed many interactions not found when analyzing the conditions individually. In particular, the comparison revealed many interactions between DNA damage-response genes, which suggests that the proteins from these gene work together in mechanisms that arise in the MMS condition. When the gene interactions were mapped onto protein interactions, the researchers identified the differentially active connections as connecting between protein complexes. They interpreted the protein complexes as stable structures that are differentially recruited into mechanisms when specific tasks need to be performed.

In machines, components that are not needed on a given occasion just sit idle. Given the ability of organisms to synthesize proteins as needed, it is not surprising that, unlike machines, biological mechanisms construct mechanisms as need and then degrade them when not needed.

3. Mechanism 2.0

3.1. Recap

We have characterized Mechanism 1.0 in terms of the examples that were advanced in philosophical discussions of mechanistic explanation, and the features highlighted in discussions of them—such as the types of parts organization employed, and the kinds of discovery strategies used to identify such mechanisms. Our aim has not been to challenge the explicit definitions of mechanism that advocates of mechanistic explanation have advanced. Our justification is that it was the examples themselves and the discussions that emerged from considering them that made mechanistic explanation appear as a compelling alternative to other accounts of explanation, and that informed the community's understanding of the contrast between mechanistic and other forms of explanation.

Pursuing the examples advanced for Mechanism 1.0 has been extremely fruitful. The examples had the virtue of being widely intelligible by philosophers without extensive background in biology. And they painted a sharp and, we think, justifiable, contrast with alternative forms of explanation, including law-based explanations common in physics and some other sciences. However, as we have tried to show, these examples do not reflect the full scope of mechanistic accounts offered in science. The examples presented a view that is simplified along several key dimensions: portraying mechanisms as having discrete and enduring parts, organized in relatively fixed ways, and clearly distinguished from the environment in which they operated.

We have put forward other examples that differ in important respects. In many the operation of the mechanism depends on concentrations, rather than discrete parts. In some cases the parts change over time. Moreover, we presented mechanisms whose organization changes as the mechanism functions and which bled into their environment (including other mechanisms) rather than being sharply distinguished from it. Finally, we identified mechanisms that are transient, not enduring. All told, this puts the machine image underlying Mechanism 1.0 under severe strain.

3.2. Moving beyond the traditional machine metaphor

The examples put forward on behalf of Mechanism 1.0 conformed closely to the picture of machines designed by humans. Thinking about them as machines facilitated discovering and reasoning about them. Like machines, they are assumed to be localized in their environments and to have parts that are identified in terms of their structure. Researchers expect to be able to trace activity through the mechanism as it generates the phenomenon. Their internal states may change in the course of processing, but the overall organization remains constant. Several mechanists sought to differentiate biological mechanisms from

machines, but given the examples advanced to support Mechanism 1.0, the difference between biological mechanisms and machines was not all that clear.

The departures from Mechanism 1.0 on which we have focused put even more strain on the machine metaphor. We need to recognize, however, that *machine* is also an evolving notion. Historically, those opposing mechanism—vitalists and holists—emphasized the differences between biological systems and the machines then present (Bechtel, 2016). Inspired in part by Bernard and Cannon, cyberneticists (Wiener, 1948) expanded on historical conceptions of machines by emphasizing such things as the potential for control provided by negative feedback. Negative feedback, and its capacity for facilitating not only control but also oscillatory behavior, was a major inspiration for Mechanism 1.1. Sustained oscillators that can synchronize with each other and be entrained to external oscillations, provided examples of systems that don't wait for input to initiate activity but are endogenously active.

Today our conception of machines continues to evolve as designers explore options to make physical devices behave in ways not conceived of in the past. The ability to control electrical activity in computers through software, including software that can be modified by the machine itself, has certainly fostered this expansion in the concept of a machine. An additional factor has been the application of ideas about organization discovered in biology to designing new machines. Some of these are ideas we have characterized as motivation for Mechanism 2.0; incorporating them into machines may once again reduce the gap between machines and mechanisms.

Regardless of whether machines continue to evolve to more resemble biological mechanisms, our focus is on how biologists are revising their conception of how mechanisms are structured and, especially, of how mechanistic explanations work. Expectations for a localized, graphically depicted system, whose workings can be tracked via mental rehearsal and where discovery consists in large part of "looking under the hood," have been altered. Many biologists now recognize that mechanisms that differ from those advanced for Mechanism 1.0 are both common and important, and that modeling them requires advanced, typically mathematical, methods that go beyond flowcharts and structural figures.

3.1. Next steps

Where does this leave us? First, insofar as we have focused on the examples offered, not the definitions advanced, we are not contesting the definitions of mechanism. The letter of these definitions may well be compatible with several of the examples we have advanced. We don't see much benefit in the project of defining mechanism. Glennan (in press) advances the notion of *minimal mechanism* as providing a common basis for various more specific conceptions of mechanism, and this may suffice. Speaking loosely, we can treat any explanation that appeals to underlying parts-and-organization as mechanistic. This rough way of pointing to the kinds of systems and explanations at issue is all that we need.

If we are not contesting the definitions, then why are we making so much out of the ways our examples differ from those we see as characteristic of Mechanism 1.0? For one thing, we think it is the examples and the way they have been discussed that have shaped our understanding of mechanistic explanation. But more crucially, the interest in mechanistic explanation is not focused on developing adequate definitions but on issues such as how mechanistic accounts explain, how they are discovered and evaluated, and the ways in which they get applied in scientific reasoning. On this score, the examples fitting Mechanism 1.0 pointed in particular directions—the focus was on a well-delineated set of entities that were organized in a stable manner, were demarcated from others, and endured. With these examples in mind, it was natural to pursue particular kinds of strategies of research. For instance, decomposition and localization—i.e. breaking down a system into its parts and identifying their structure and place in the mechanism's layout—were emphasized by Bechtel and Richardson (1993). Strategies for constructing mechanistic hypotheses also followed this pattern, For instance, in the examples illustrating Darden's (2006) and Craver and Darden's (2013) strategy of modular subassembly (i.e. hypothesizing that a mechanism is composed of modules known from other mechanisms, in an altered configuration), the parts are presented as discrete and only interacting via their inputs and outputs. In their strategy of forward/backward chaining, one uses information about early (or, correspondingly, late) stages in the mechanism's operation to sketch hypotheses about later (or, correspondingly, earlier) stages. It is unclear that such a strategy can succeed when the mechanisms does not have well-defined boundaries, so that the kinds of constraints placed by earlier stages on the process are very hard to pin down. Thus, these and similar research strategies largely rely on discrete and localized parts, sequential organization and well-demarcated boundaries. They are unlikely to succeed with the types of mechanisms we have drawn attention to.

The mechanisms we have advanced on behalf of Mechanism 2.0 require different approaches. In cases in which concentrations matter, researchers often need to measure concentrations and collect time series data to understand how they change. If the mechanism only exists in certain contexts or if organization changes in different circumstances, then researchers need to contextualize the study of the mechanism and use tools that allow them to discern which components and processes are active when and where. Differential network biology, briefly discussed above, represents one such strategy. One should expect that the parts and organization of a mechanism may change as conditions in which it operates change. If mechanisms are not sharply differentiated but bleed out into their environment, then researchers may need to make different choices about the identity of the mechanism on different occasions.

As we are not advocating definitions, in advancing the notion of Mechanism 2.0 we are not seeking new definitions. Rather, we are advocating a broader research agenda that seeks different ways in which mechanisms can depart from the examples advanced for Mechanism 1.0 and still count as mechanisms. The more important task is likely to be distinguishing different kinds of mechanisms, potentially generating a taxonomy of

mechanisms. One dimension in such a taxonomy might be whether parts are localized or not and whether it is the structure of the parts or their concentrations that matter. Another might be whether the organization is enduring, changes under endogenous control, or changes in different environments. A taxonomy is only valuable if the different types of mechanism that are distinguished matter for philosophical and/or scientific objectives. We have made some suggestions as to how the different departures from Mechanism 1.0 do matter for understanding how a mechanism serves to explain a phenomenon and how the mechanism is discovered.

A taxonomy of mechanisms would address the diversity of mechanism types. A further set of questions concern mechanism tokens, namely: what are the identity conditions for mechanisms. When can we say that the same mechanism has changed over time and when do we have a new mechanism? Under Mechanism 1.0 this question hardly ever arose, but once we have mechanisms with changing parts, shifting organization and fuzzy or non-existent boundaries, it is natural to wonder whether and when one can speak of a mechanism as persisting through time.

We cannot resolve this issue here, of course, and it may very well be that the answer can only be given as part of a broader story about objects, change and identity through time. But let us indicate three potential directions on might proceed to individuate mechanisms. First, one can move to an "ephemeral mechanism" outlook, i.e. accept that when components and/or organizational features change, as they often do, we no longer have the same underlying mechanism. This entails that one and the same phenomenon may often be underpinned by different mechanisms over time. This is a somewhat unintuitive idea, as it severs, or at least weakens, the link between mechanisms and phenomena. But perhaps it is correct and unproblematic. A second option is to identify mechanisms via the phenomena they explain. On this approach, so long as we have the same phenomenon we have the same mechanism. Of course, this raises questions about the identity of phenomena. Not much has been written on this. The only well-developed account, Kaiser and Krickel (2016), construes phenomena as "object-involving occurrences." As the name suggests, this account presupposes a notion of (biological) objects, and it is not clear that such a notion can be retained in light of the cases we have looked at. Finally, one can view Mechanism 2.0 in the context of process ontology, the idea that the biological world, perhaps the world at large, consists of processes rather than objects (i.e. a temporally extended, constantly changing, "stream" rather than stably structured "thing"). Dupré (2012, 2014) has recently been arguing for such a view, and the cases we have discussed may provide more grist for his processual mill. Some will find process ontology to be a radical and implausible viewpoint, although there are ways of making it compatible with an explanatory appeal to mechanisms, especially if, as we tend to think, explanations are to be seen in epistemic rather than ontic terms. Here we remain uncommitted, as our principal aim is to highlight the issue as a topic for further exploration.

In concluding, we should stress again why we treat the examples we have described as mechanistic explanations despite the fact that some would treat them as non-mechanistic.

We think it is both more useful philosophically and in better accord with scientific practice, to expand our perspective on the types of mechanism that occur in biology. In expanding the scope of what qualifies as a mechanism, however, we are not emptying the notion *mechanism* of content. For one thing, the contrast between mechanistic explanation and DN or other formalist views of explanation is retained. Explanation is still a matter of describing the causal underpinnings of a phenomenon, rather than embedding it in a formal deduction schema. But beyond that, there are several sorts of explanations that can be seen to be non-mechanistic: etiological explanations, which chart a causal process leading up to an event are one example, as well as teleological explanations, which describe the function of an object or feature. Arguably, so are mathematical explanations (Pincock, 2007; Lange, 2012), which account for a phenomenon in terms of a formal-mathematical properties instantiated by it.

Clearly, we have only begun the task of transitioning from Mechanism 1.0 to 2.0. If our suggestion that there are multiple dimensions in which recognizably mechanistic science departs from Mechanism 1.0, Mechanism 2.0 might not have a univocal characterization but offer a taxonomy. As with the initial articulation of mechanism 1.0, we expect this project to be driven by close attention to the explanations actually offered in science. We also expect it to be as fruitful philosophically. And, if successful, it might ultimately itself be found wanting, making way for Mechanism 3.0.

References

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M. K., Chuang, R., Jaehnig, E. J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., Fiedler, D., Dutkowski, J., Guenole, A., van Attikum, H., Shokat, K. M., Kolodner, R. D., Huh, W. K., Aebersold, R., Keogh, M. C., Krogan, N. J., & Ideker, T. (2010). Rewiring of genetic networks in response to DNA damage. *Science, 330*, 1385-1389.

Bass, J., & Takahashi, J. S. (2010). Circadian integration of metabolism and energetics. *Science, 330*, 1349-1354.

Bechtel, W. (2006). *Discovering cell mechanisms: The creation of modern cell biology*. Cambridge: Cambridge University Press.

Bechtel, W. (2015). Can mechanistic explanation be reconciled with scale-free constitution and dynamics? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*.

Bechtel, W. (2016). Mechanists must be holists too! Perspectives from circadian biology. *Journal of the History of Biology*, 1-27.

Bechtel, W., & Abrahamsen, A. (2010). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science Part A, 41*, 321-333.

Bechtel, W., & Richardson, R. C. (1993/2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.

Brigandt, I. (2013). Systems biology and the integration of mechanistic explanation and mathematical explanation. *Studies in History and Philosophy of Biological and Biomedical Sciences, 44*, 477-492.

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. Chicago: University of Chicago Press.

Darden, L. (2006). *Reasoning in biological discoveries: Essays on mechanisms, interfield relations, and anomaly resolution*. Cambridge: Cambridge University Press.

de Lichtenberg, U., Jensen, L. J., Brunak, S., & Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science, 307*, 724-727.

Dupré, J. (2012). *Processes of life: Essays in the philosophy of biology*. Oxford ; New York: Oxford University Press.

Dupré, J. (2014). A process ontology for biology. *The Philosophers' Magazine, 67*, 81-88.

Glennan, S. (in press). *The new mechanical philosophy*. Oxford: Oxford University Press.

Ideker, T., & Krogan, Nevan J. (2012). Differential network biology. *Molecular Systems Biology, 8*, 565.

Kaiser, M. I., & Krickel, B. (2016). The metaphysics of constitutive mechanistic phenomena. *The British Journal for the Philosophy of Science*.

Lange, M. (2012). What makes a scientific explanation distinctively mathematical? *The British Journal for the Philosophy of Science*.

Levy, A. (2013). What was Hodgkin and Huxley's Achievement? *The British Journal for the Philosophy of Science*.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1-25.

Marom, S. (2010). Neural timescales or lack thereof. *Progress in Neurobiology, 90*, 16-28.

Moore, R. Y., & Eichler, V. B. (1972). Loss of a circadian adrenal corticosterone rhythm following suprachiasmatic lesions in the rat. *Brain Research, 42*, 201-206.

Nathan, M. J. (2014). Causation by concentration. *British Journal for the Philosophy of Science, 65*, 191-212.

Nelson, P. C. (2015). *Physical models of living systems*. New York, NY: W.H. Freeman & Company.

Pincock, C. (2007). A role for mathematics in the physical sciences. *Nous, 41*, 253-275.

Reppert, S. M., & Weaver, D. R. (2001). Molecular analyses of mammalian circadian rhythms. *Annual Review of Physiology, 63*, 647-676.

Rust, M. J., Markson, J. S., Lane, W. S., Fisher, D. S., & O'Shea, E. K. (2007). Ordered phosphorylation governs oscillation of a three-protein circadian clock. *Science, 318*, 809-812.

Sahar, S., & Sassone-Corsi, P. (2009). Metabolism and cancer: the circadian clock connection. *Nature Reviews Cancer, 9*, 886-896.

Tseng, R., Chang, Y. G., Bravo, I., Latham, R., Chaudhary, A., Kuo, N. W., & LiWang, A. (2014). Cooperative KaiA-KaiB-KaiC interactions affect KaiB/SasA competition in the circadian clock of Cyanobacteria. *Journal of Molecular Biology, 426*, 389-402.

Welsh, D. K., Logothetis, D. E., Meister, M., & Reppert, S. M. (1995). Individual neurons dissociated from rat suprachiasmatic nucleus express independently phased circadian firing rhythms. *Neuron, 14*, 697-706.

Welsh, D. K., Takahashi, J. S., & Kay, S. A. (2010). Suprachiasmatic nucleus: Cell autonomy and network properties. *Annual Review of Physiology, 72*.

Wiener, N. (1948). *Cybernetics: Or, control and communication in the animal and the machine*. New York: Wiley.

Zhang, E. E., Liu, A. C., Hirota, T., Miraglia, L. J., Welch, G., Pongsawakul, P. Y., Liu, X., Atwood, A., Huss, J. W., Janes, J., Su, A. I., Hogenesch, J. B., & Kay, S. A. (2009). A genome-wide RNAi screen for modifiers of the circadian clock in human cells. *Cell, 139*, 199-210.

A Pursuit Worthiness Account of Analogies in Science

Abstract: Analogies often provide reasons for pursuing hypotheses or models. This is illustrated with a case study on the liquid drop model of the atomic nucleus. I criticise accounts in which analogies provide reasons for pursuit through epistemic support, proposing instead that analogies increase the value of learning the truth. I consider two accounts of this type: first, that analogies indicate potentials for theoretical unification; second, that analogies facilitate the transfer of already well-understood modelling frameworks to new domains. While the first is plausible for some cases, only the second can account for the liquid drop case study.

1. Introduction

For much of the 20th century it was hotly contended whether analogies play any normatively interesting role in scientific reasoning. Defending analogies, Norman Campbell (1920, ch. 6) and Mary Hesse (1966) responded to Pierre Duhem (1914/1954) and his intellectual heirs among the logical empiricists, such as Hans Reichenbach. Although the latter critics admitted (grudgingly) that analogies sometimes guide the development of scientific theories, they regarded this as a mere psychological curiosity, not something that plays any interesting normative role in scientific reasoning (e.g. Reichenbach 1944, 66-72). Arguing that analogies serve important purposes that philosophers of science ought to account for, Campbell and Hesse opposed these at the time widely accepted views.

Today, most philosophers interested in the issue agree that analogies play an important role in scientific reasoning. A number of different roles for analogies have been discussed (Bartha 2013, §1). Some challenge the presumption that generative reasoning is beyond the scope of normative theorising. For instance, Nersessian (1988), drawing on cognitive psychology and computational modelling, has argued that analogies can function as heuristics for developing or articulating scientific theories in ways that are both "systematic and subject to evaluation" (1988, 42). Call these *generative accounts* of analogical reasoning. Others take analogies to provide epistemic support for hypotheses and consequently propose accounts of how or when analogical arguments can provide this kind of support. Call these *justificatory accounts*.

My focus in this paper is on what can be called *pursuit worthiness accounts*, i.e. accounts according to which analogies provide reasons for testing or developing a hypothesis further.[1] While compatible with the other two, pursuit worthiness accounts are necessary for explaining some aspects of scientific reasoning that cannot be captured by purely justificatory or generative accounts. To illustrate this point, I outline a case study in Section 2, involving the early development of the liquid drop model of the atomic nucleus. I argue that in this case the liquid drop analogy motivated physicists to pursue the model despite it initially facing empirical and theoretical problems. In the remainder of the paper I consider different accounts of how analogies justify pursuit.

I start, in Section 3, by criticising accounts defended by Wesley Salmon (1967) and Paul Bartha (2010), according to which analogies provide reasons for pursuing a hypothesis in virtue of providing reasons for their truth. I argue that even if analogies sometimes provide epistemic support, this is not always a reason in favour of pursuit. Instead, I propose that analogies are better seen as justifying pursuit by increasing the value of learning whether the hypothesis is true. In Section 4 I consider an account where hypotheses based on analogies have a high potential for unification. I argue that while this account is plausible for some cases, it does not fit the case of the liquid drop model. Finally, in Section 5, I propose an alternative account of this case according to which analogies facilitate the transfer of an already well-understood modelling framework to a new domain of phenomena.

2. Case Study: The Development of the Liquid Drop Model

The liquid drop model of the atomic nucleus was developed from the late 1920s onwards, during a time where physicists were trying to extend their understanding of the structure of atoms to the atomic nucleus itself.[2] The model was first proposed in 1928-29 by George Gamow, at the time a Russian doctoral student visiting Western Europe, who suggested

[1] I borrow the terminology of 'generative', 'justificatory' and 'pusuitworthiness' accounts from McKaughan (2008).

[2] The following is based on Stuewer's (1994) account.

that the nucleus "may be treated somewhat as a small drop of water in which the particles are held together by surface tension" (cited from Stuewer 1994, 80). In line with common assumptions at the time, he modelled the nucleus as consisting of a collection of α-particles, and assumed that the nucleus was in equilibrium between the kinetic energy of the particles and the surface tension. On this basis Gamow then tried to derive an expression for the mass defects (i.e. the nuclear binding energy) of the different nuclei.

Niels Bohr and Ernest Rutherford were enthusiastic about the model, providing support for Gamow to develop it from 1929 to 1931. However, while Gamow made some progress, he quickly ran into problems. Although his theoretically predicted mass defects traced a curve of the same general shape as the experimentally determined ones, it only gave reasonably accurate quantitative predictions for the lighter elements. He suspected this could be remedied by taking into account the nuclear electrons that were thought to exist at the time. However, when he tried to incorporate these into his model he ran into a major theoretical problem (the so-called Klein paradox) that he was unable to overcome. Consequently, by the summer of 1930 Gamow began to turn his attention elsewhere (Stuewer 1994, 78-85).

Despite these problems, the model quickly became popular among physicists, not because they were confident it accurately represented the nucleus, but as a speculative attempt to solve certain problems. For instance, in 1930 Rutherford wrote that the model "while admittedly imperfect and speculative in character is of much interest as the first attempt to give an interpretation of the mass-defect curve of the elements" (cited from Steuwer 1994, 86-7). During the 1930s the model was further developed, following two broad trajectories. First, following the discovery of neutrons in 1932, Werner Heisenberg and Carl von Weizsäcker tried to revise the assumptions of the model to yield an empirically more accurate mass defect curve (*ibid*., 87-97). Second, Bohr and several others modified the model in order to account for artificially induced radioactivity (i.e. radioactive elements produced by bombarding stable elements with neutrons) as an excitation and subsequent 'evaporation' of particles from the drop of 'nuclear fluid' (97-

107).[3] Finally, in 1938-39 Lise Meitner and Otto Frisch, combining insights from both research programmes, realised that the liquid drop model could be adapted to explain nuclear fission, a newly discovered and at the time highly puzzling phenomenon (107-116).[4]

As is clear from the latter part of this story, the analogy played an important role in guiding the revisions and extensions of Gamow's original model. This use of analogy is what generative accounts aim to analyse. I return to this use of the liquid drop analogy in Section 5. For now, I want to highlight that already when Gamow proposed the liquid drop model in 1928-30, it was received positively and was taken up by a number of physicists, despite its initial problems. The analogy also seems to have motivated pursuing the model in the first place, before there was any particular reason to think it even approximately true. The question that I will focus on in the rest of this paper is *why* it was more reasonable to spend time and resources pursuing this particular model, rather than some alternative mathematical model not grounded in analogies.

3. Pursuit Worthiness, Plausibility and Probability

It might be thought that there is a straightforward answer to this question. Although there might not have been grounds for *accepting* Gamow's model in 1930, the analogy could still have shown it *plausible* and, the idea goes, the fact that the model was plausible made it reasonable to pursue it. But since reasons for regarding a model or hypothesis as plausible are merely a weaker form of epistemic support, these are not fundamentally different from reasons for its truth.[5]

A version of this account was suggested by Wesley Salmon (1967). Salmon was responding to N.R. Hanson's (1958, 1074) claim that there is a fundamental difference between reasons for accepting a hypothesis H and "reasons for suggesting H in the first

[3] A number of alternative (but sometimes related) analogies also influenced this line of physical theorising about atomic nuclei (Stuewer, *ibid.*).

[4] See also Andersen (1997) on the experimental and theoretical developments which lead to the discovery of fission.

[5] See Kordig (1978) for an account along these lines, not specifically concerned with analogies.

place" since the latter are "reasons which make H a *plausible conjecture*" (*ibid.*). Hanson (1077-79) argued that reasons for suggesting hypotheses (what I here call reasons for pursuit) can be based on analogies, among other things. Against this, Salmon (1967, 113-18) argues that plausibility judgements should be understood as estimates of the prior probability of a hypothesis. Since in a Bayesian framework it is necessary to make some judgement of prior probabilities to evaluate the posterior probability of a hypothesis, this furnishes an important role for plausibility judgements without these being fundamentally different from reasons for acceptance. According to Salmon, analogical arguments are plausibility arguments in this sense (127).

Whereas Salmon thus equates reasons for pursuit with estimates of prior probability, Paul Bartha's (2010) recent work on analogical reasoning gives a more nuanced account of their relation. I here outline some details of Bartha's account of analogical reasoning, since I draw on some of them later on. Following Hesse (1966, 59), Bartha endorses a *two-dimensional analysis* of analogical arguments. While many accounts only focus on *horizontal relations*, i.e. the similarities and differences between the source and target system, two-dimensional accounts also emphasise the *vertical relations*, consisting of dependency relations (e.g. causal, modal or explanatory relations) within the two domains. Building on this idea, Bartha (2010, ch. 4) defends an inference schema that may be summarised as follows:

(B1) There is some structure of dependency relations *R*(a, b, c, …) between features a, b, c, … of the source system, S1. [*Prior association*].

(B2) The target system, S2, has one or more features a', b', c', … analogous to a, b, c, … [*Potential for generalisation*].

(B3) S2 does not have any features which would preclude *R*' (analogous to R) from obtaining. [*No critical difference*].

Therefore:

(B4) It is prima facie plausible that *R*'(a', b', c', …) obtains for S2, and *a fortiori* that S2 has features a', b', c', …

The first premise states that there is a "prior association" in S1, in the form of some structure of dependency relations between its features. Which kinds of dependency relations to look for varies between contexts, but a good example of a structure of dependency relations is how the parts of a mechanism interact and constrain each other to produce certain effects. Second, we look at whether there is a "potential for generalisation", meaning that the target system has some features analogous to those involved in the prior association in S1. Finally, we consider whether there are any "critical differences" between the two systems, i.e. whether S2 has any features precluding a relation analogous to the prior association from obtaining. Given these premises, according to Bartha, it is prima facie plausible to "transfer" the prior association to the target system, and thus infer that the relevant further features involved in the prior association obtain in S2 as well.

Bartha highlights that arguments of this type are often used to support hypotheses before they have been tested (2010, 6) and that they provide reasons for investigating hypotheses further (16). Like Salmon, he thinks this is because analogies support plausibility judgements, but Bartha does not equate plausibility judgements with estimates of prior probability. That a hypothesis *p* is 'prima facie plausible', he takes instead to mean "roughly speaking, … There are sufficient grounds for taking *p* seriously" (2010, 16). This is partly an epistemic notion. A plausible hypothesis, according to Bartha, "has epistemic support: we have some reason to believe it, even prior to testing" (15) and it has "an appreciable likelihood of being true" (18). But he also takes plausibility judgements to have pragmatic connotations: "To say that a hypothesis is plausible typically implies that we have good reason to investigate it (subject to the feasibility and value of investigation)" (15). In a suggestive footnote (p. 18, note 19) Bartha furthermore mentions that reasons for pursuit depend on epistemic support "in a decision-theoretic sense" given "contextual information about costs and benefits." However, he adds that absent this information "the two points are at least partially independent" (*ibid.*). So although epistemic support is important to what Bartha means by plausibility, considerations about 'feasibility' and 'value' are relevant as well.

Given this elucidation of what he means by 'prima facie plausibility', it is

consistent with Bartha's account that analogical inferences can provide reasons for investigating a hypothesis without necessarily providing reasons for its truth. However, in practice he tends to focus on epistemic support. For instance, he claims, "Any argument that a hypothesis is prima facie plausible … should provide reasons to think the hypothesis might be true" (18). Furthermore, he still follows Salmon in identifying a hypothesis' *degree* of plausibility with its prior probability (e.g. pp. 15-6, 291-302). As I read Bartha, analogies primarily provide reasons for pursuing hypotheses by providing epistemic support for them. Once this is established, whether we are then justified in pursuing a hypothesis all things considered depends on 'contextual information', i.e. information in addition to that provided by the analogy, about the costs and benefits of pursuing it.

Although Salmon and Bartha might be right that analogies sometimes give reasons for ascribing higher prior probability to a hypothesis, I do not think this gives a satisfactory account of how analogies justify pursuit in cases like the liquid drop model. First, it is not clear that physicists in 1930 regarded the liquid drop model as significantly more probable than so many other possible models. Second, while I agree with Bartha that having reasons for pursuing a hypothesis can be elucidated in decision-theoretic terms, he fails to take the implications of doing so fully into account. Since being justified in pursuing a hypothesis depends on a number of factors apart from its epistemic support, why assume that the analogy increased its epistemic support rather than some of the other factors? One cannot simply assume that when analogies motivate pursuing a hypothesis, the analogy must therefore have provided reasons for its truth. Third, it is not always the case that increasing the probability of a hypothesis is a reason in *favour* of pursuing it, let alone a sufficient reason.

When considering whether to pursue a hypothesis H, we need to take into account the different possible outcomes of doing so. We might learn that H is true, but we might equally learn that it is false. Furthermore, we should also take into account the possibility of getting no useful evidence or – even worse – getting misleading evidence, i.e. evidence that leads us to mistakenly accept or reject H. Following Nyrup (2015, 755-6), this can be represented in a simple decision-theoretic model. Suppose we only distinguish between two possible states of the world, that H is true and that it is false, and that we are interested

in a range of epistemic attitudes EA_1, $EA_{2, \dots,}$ EA_n we might end up having towards *H*, (e.g. accepting *H*, rejecting *H* and staying agnostic).[6] Then the expected utility of pursuing *H* is given by:

$$
\begin{aligned}
(1) \quad EU(p(H)) \quad &= \quad Pr(H) \times \sum [U(EA_i(H), H) \times Pr(EA_i(H) \mid p(H), H)] \\
&+ \quad Pr(\neg H) \times \sum [U(EA_i(H), \neg H) \times Pr(EA_i(H) \mid p(H), \neg H)] \\
&- \quad C(p(H))
\end{aligned}
$$

The unconditional probabilities in this model represent the probability of *H* being true (or false, respectively) at the given state of inquiry, before further testing. They can both be initial probabilities prior to *all* testing or posterior probabilities given previous testing in situations where we are considering whether to pursue *H* further. It is this quantity that Salmon and Bartha take analogical arguments to manipulate. The conditional probabilities represent how likely we are, given that *H* is true (or false), to obtain evidence sufficient to adopt the attitude EA_i towards *H*. For instance, if EA_1 is acceptance then $Pr(EA_1(H) \mid p(H), H)$ represents how likely we are to get *reliable* evidence in favour of *H*, while $Pr(EA_1(H) \mid p(H), \neg H)$ is how likely we are to get *misleading* evidence in favour of *H*. $U(EA_1(H), H)$ represents the value of, e.g., correctly accepting *H*, while $U(EA_1(H), \neg H)$ measures how problematic it would be to mistakenly accept *H*, and mutatis mutandis for other epistemic attitudes. Finally, $C(p(H))$ is the cost (time, resources, etc.) of pursuing *H*.[7]

This analysis highlights that there are a number of different factors relevant to whether it is worth pursuing a hypothesis. In order for an argument to provide additional reasons for pursuing *H*, it must be the case that it increases our estimate of $EU(p(H))$. But there is no reason to suppose that it must increase the probability of *H* being true rather than, e.g., showing that it would be more interesting to know whether *H* is true, showing that *H* is less costly to pursue or showing that pursuing *H* is more likely to produce reliable

[6] It is possible to include further states of the world, e.g. various degrees to which *H* is partially true, or a broader range of epistemic attitudes without changing the conclusions I draw from this model.

[7] I assume for simplicity that these costs are commensurable with the utility of knowing whether *H* is true and that the costs of pursuing *H* are independent of its truth.

evidence. In fact, unlike these other factors, it is *not* generally the case that increasing Pr(*H*) raises EU(p(*H*)). For instance, if it would be easy to falsify *H* but difficult to get reliable evidence to confirm it, or if knowing that *H* is false would be more interesting than knowing that it is true, *reducing* Pr(*H*) could raise EU(p(*H*)) (cf. Nyrup 2015, 759).

4. Analogies as Guides to Unification

I have so far criticised the assumption that analogies provide reasons for pursuit by providing epistemic support. I propose that analogies instead justify pursuing *H* by increasing the value of knowing whether *H* is true. I develop this proposal in the remainder of this paper. More specifically, I consider two accounts of this type. I start with the idea that analogies indicate hypotheses that would provide increased theoretical unification, if shown true. While plausible for some cases, I will propose an alternative account in the next section which better accounts for the liquid drop case.

Campbell's defence of analogies in physics was arguably based on the unificationist idea. While he thought that theories based on mechanical analogies are more likely to be false than ones which merely posit generalised laws extrapolated from observed regularities (152), he argues that analogically based theories are valuable "simply because the ideas which they bring to mind are intrinsically valuable" (1920, 132). The reason is that they offer the chance of laws capable of unifying quantities from previously distinct domains, e.g., heat and momentum, in the case of the billiard ball model of gases. Insofar as we consider it an intrinsically valuable project to achieve this kind of unification, we "must balance that value against the chance of error" (152). Although Campbell does not elaborate much further on these remarks, it is clear that the value he ascribes to analogies is not that they provide increased epistemic support for theories.

The idea that the value of obtaining unifying theories has to be balanced against the risk of error fits the decision-theoretic model outlined above. If we agree with Campbell that it is intrinsically valuable to discover that a unifying theory is true, this would increase the first term of equation (1). If this value is sufficiently high, it could outweigh a decreased prior probability, which would otherwise shift the weight towards the second term of the equation (but notice, again, that reducing prior probability does not necessarily

decrease overall the expected utility of pursuit).

This account also fits one line of justification Bartha (2010, ch. 7) offers for his account, viz. that it tends to promote the traditional theoretical virtues, in particular unification.[8] If we construe unification as the ability to explain a wide range of phenomena using the same basic explanatory pattern (Kitcher 1989), we can see how this fits Bartha's inference schema. Premise (B1) identifies the existence of the explanatory pattern *R* (the prior association) in S1, while (B2) points out that there are a number of features in S2 that could potentially be explained by the same pattern. Since (B3) there is no known reason to rule out this possibility, there is a potential for unifying the relevant features of S1 and S2 in single explanatory schema. So if we were to discover that *R* holds for S2, we would have increased the unification of our knowledge of the world.

In my view, this account of analogies provides a plausible account of how analogical reasoning justifies pursuit in some cases but not all. In cases such as the billiard ball analogy for gases or the 'waves in a mechanical medium' analogy for light (discussed e.g. by Hesse 1966, Nersessian 1988), the analogies do seem to promise to unify thermodynamical and optical phenomena, respectively, with the theoretical framework of classical mechanics. From the perspective of nineteenth-century physicists, these analogies pointed to potential increases in theoretical unification. However, this story does not work for cases like the liquid drop model. Although Bohr, Rutherford and other physicists regarded Gamow's analogy as suggesting a very promising line of research, this does not seem to be because it promised to unify the physics of water drops and atomic nuclei. The liquid drop model employs modelling techniques analogous to those applied to water drops, but it was clear that the explanations for the two kinds of phenomena would be very different. Even if one might hope that an increased understanding of the atomic nucleus could eventually lead to a unified account of the two types of systems, the liquid drop

[8] Bartha argues that analogies are also conducive to coherence, simplicity and fruitfulness, but regards unification as the most central. Bartha (2010, 256) here recognises that as long as we consider these virtues valuable to achieve, this is sufficient to show a hypothesis 'plausible' in his sense of 'worthy of investigation'. However, he also suggests that his argument can be combined with the argument that the theoretical virtues are "indicators of empirical adequacy (or truth)" (*ibid.*).

model does not in itself promise to achieve this kind of unification in the way that the billiard ball model and mechanical ether models did.

5. Transferring Modelling Frameworks Through Analogies

In order to account for how analogies justify pursuit in cases like the liquid drop model, we need to switch to a more dynamic account of the relation between analogies and scientific models. I have so far focused on whether analogies can justify pursuing a specific hypothesis. However, in the liquid drop case, Gamow and those who subsequently worked on the liquid drop model did not exactly pursue any specific hypothesis about the structure of the atomic nucleus. Rather, they tried to model the atomic nucleus as if it were a water drop in order to construct a potential explanation of some otherwise puzzling phenomenon – i.e. the mass defect curve for Gamow, Heisenberg and Weizsäcker, artificial radioactivity for Bohr and his colleagues, and nuclear fission for Meitner and Frisch. They were of course still interested in achieving a correct (or at least empirically accurate) description of the nucleus, but their first priority was to formulate a potential explanation of the target phenomenon. Rather than pursuing a specific *hypothesis*, the water drop analogy motivated the pursuit of the *research project* of adapting a modelling framework to the atomic nucleus for certain explanatory purposes. Or, if we want to say that they pursued a hypothesis, it was not one of the form "the atomic nucleus has features a, b, c, … analogous to a water drop" but rather something like "modelling the atomic nucleus analogously to a water drop can provide a (correct) explanation of phenomena x, y, z, …."

That analogies guide the development hypotheses is also emphasised by proponents of generative accounts, such as Nersessian (1988). But it is important to notice that adopting a dynamic view of the relation between models and analogies does not in itself answer the question of why it was reasonable to pursue an analogical modelling framework, rather than so many others. This is how pursuit worthiness accounts differ from generative accounts. The latter primarily describe the cognitive role analogies play in shaping and guiding the development of novel scientific theories. Pursuit worthiness accounts, by contrast, justify why one should choose to develop a theory using analogies in the first place.

That analogies should be a *help* in developing theories is not obvious. Campbell (1920, 130), for instance, disagreed: "Analogy, so far from being a help to the establishment of theories, is the greatest hindrance. It is never difficult to find a theory which will explain the laws logically; what is difficult is to find one which will explain them logically and at the same time display the requisite analogy. … To regard analogy as an aid to the invention of theories is as absurd as to regard melody as an aid to the composition of sonatas." Now, *pace* Campbell, it might be that imposing constraints actually makes it easier to come up with genuinely novel ideas. However, the core point here is that the relevant question is not how to most effectively come up with *novel* ideas, but rather how to come up with *ideas that are worth pursuing*. Sometimes, e.g. if we lack any possible explanations, coming up with genuinely novel ideas might be intrinsically desirable. But in other cases, e.g. if we are overwhelmed by too many hypotheses, we may instead prefer to *restrict* ourselves to generating hypotheses of high quality.

So why are modelling frameworks based on analogies more pursuit worthy in cases like the liquid drop model? I want to end by proposing that these frameworks are more pursuit worthy because they facilitate the transfer of a modelling framework to construct explanations in a new domain.[9] One simple reason for trying to adapt an already existing modelling framework to a new case is that this is typically easier and less time consuming than developing a new one from scratch. Thus, transferring a modelling framework by analogy can often reduce the costs of pursuit.

However, constructing new explanations using analogically transferred modelling frameworks arguably also increases the potential understanding one can achieve through those explanations. This is because achieving scientific understanding of some phenomenon depends upon having a well-understood modelling framework. Understanding *why* a phenomenon occurs requires that one understands the model one

[9] This account is inspired by Hesse's and Bartha's idea that analogical inferences "transfer" explanations from one domain to another. However, as emphasised above, I focus on adapting a framework to produce new explanations rather than one-off inferences. In this respect, it is closer to Hesse's (1966: 157-177) suggestion that analogies provide a form of explanation by "metaphorically redescribing" the target domain in terms of the source analogy.

understands the phenomenon *with* (Strevens 2013: 513; cf. de Regt 2009). Thus, if an already well-understood modelling framework can be adapted to produce a correct explanation, little work is needed to realise its explanatory potential. One might eventually achieve a similar understanding of a new, purpose-built modelling framework. But, first, it would typically require a lot more effort to achieve this level of understanding. And, second, the analogically based framework offers an already proven explanatory power, as opposed to a merely potentially achievable one. In this way, even though in physicists in 1930 did not know that Gamow's model could be adapted to explain the respective phenomena they sought to explain, they still had good reasons to pursue the modelling approach indicated by the liquid drop analogy.

References

Andersen, Hanne. 1997. "Categorization, Anomalies and the Discovery of Nuclear Fission", *Studies in the History and Philosophy of Modern Physics* 27:463-492.

Bartha, Paul. 2010. *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments*, New York: Oxford University Press.

———. 2013. "Analogy and Analogical Reasoning". In *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), ed. Edward Zalta, http://plato.stanford.edu/archives/fall2013/entries/reasoning-analogy/.

Campbell, Norman. 1920. *Physics: The Elements*, Cambridge: Cambridge University Press.

de Regt, Henk. 2009. "The Epistemic Value of Understanding", *Philosophy of Science* 76:585-97.

Duhem, Pierre. 1914/1954. *The Aim and Structure of Physical Theory*. Repr. Princeton, NJ: Princeton University Press.

Hanson, Norwood Russell. 1958. "The Logic of Discovery", *Journal of Philosophy* 55:1073-1089.

Hesse, Mary. 1966. *Models and Analogies in Science*, Notre Dame: University of Notre Dame Press.

Kitcher, Philip. 1989. "Explanatory Unification and the Causal Structure of the World." In

Scientific Explanation, *Minnesota Studies in the Philosophy of Science*, vol. 13, eds. Philip Kitcher and Wesley Salmon, 410–505, Minneapolis: University of Minnesota Press.

Kordig, Carl. 1978. "Discovery and Justification", *Philosophy of Science* 45:110-117.

McKaughan, Daniel. 2008. "From Ugly Duckling to Swan: C. S. Peirce, Abduction, and the Pursuit of Scientific Theories", *Transactions of the Charles S. Peirce Society* 44:446-468.

Nersessian, Nancy. 1988. "Reasoning from Imagery and Analogy in Scientific Concept Formation". In *PSA 1988, Volume One: Contributed Papers*, eds. Arthur Fine and Jarrett Leplin, 41-47, East Lansing: Philosophy of Science Assocation.

Nyrup, Rune. 2015. "How Explanatory Reasoning Justifies Pursuit: A Peircean View of IBE", *Philosophy of Science* 82:749-760.

Reichenbach, Hans. 1944. *Philosophic Foundations of Quantum Mechanics*, Berkeley and Los Angeles: University of California Press.

Salmon, Wesley. 1967. *The Foundations of Scientific Inference*, Pittsburgh: University of Pittsburgh Press.

Strevens, Michael. 2013. "No understanding without explanation", *Studies in History and Philosophy of Science* 44:510-515.

Stuewer, Roger. 1994. "The Origin of the Liquid-Drop Model and the Interpretation of Nuclear Fission", *Perspectives on Science* 2:76-129.

What, when and how do rational analysis models explain?

(Word count: 4995)

Abstract

Probabilistic modeling is a highly influential method of theorizing in cognitive science. Rational analysis is an account of how probabilistic modeling can be used to construct non-mechanistic but self-standing explanatory models of the mind. In this article, I disentangle and assess several possible explanatory contributions which could be attributed to rational analysis. Although existing models suffer from evidential problems that question their explanatory power, I argue that rational analysis modeling can complement mechanistic theorizing by providing models of environmental affordances.

1. Introduction

During the past two decades, probabilistic modeling has become one of the most visible strands of cognitive modeling alongside connectionism, rule-based approaches and dynamical systems modeling. Curiously, against the general trend in the cognitive sciences where theorizing is increasingly anchored in neuroscience findings, probabilistic modeling of higher cognition has been a characteristically top-down endeavor. Without making any substantial commitments about the underlying cognitive mechanisms, probabilistic modeling has been applied to complex aspects of human cognition, which have often been thought of as being beyond the reach of mechanistic research methods. Models of human memory, categorization, causal learning, concept learning, and conditional inference, to mention a few applications, often show an impressive fit with empirical

data, and the novel analyses of cognitive capacities provided by the models appear to have shed new light on the nature of the explananda under study.

However, how does that shedding light actually occur – how do such computational probabilistic models explain? Although probabilistic modeling, in principle, does not rely on any particular method of explanation, modelers often refer to the idea of rational analysis as the account of how and why their models help us understand the mind (Anderson 1990; Oaksford & Chater 2007). The striking claim made by rational analysis (RA) modelers is that by understanding higher cognitive capacities as forms of inductive inference, we can predict behavior, and understand a lot about human cognition without making any assumptions about the underlying representations and processes. This agnosticism about neural and cognitive mechanisms is justified by making reference to the rationality of human behavior: We know that human agents tend to be generally well-adapted to their environment, and hence a careful analysis of the cognitive task encountered by the mind, coupled with an assumption of the optimality of human behavior, results in a putatively powerful methodology of prediction and explanation.

However, there is a large consensus in the philosophy of science that explanations also in the cognitive sciences should track causal mechanisms, and the way RA purports to sidestep the evidential and explanatory problems arising from the causal complexity of cognition has given rise to a strongly polarized debate (see, e.g., peer commentary in Jones & Love 2011). On the one hand, the way that the new mathematical methods in probabilistic modeling can combine structure and learning in human thought has lead to an exciting new paradigm for theorizing about the mind. On the other hand, the proponents of non-causal explanation need to show when and how it is that non-causal models explain rather than redescribe or merely formally unify various phenomena (cf.

Colombo & Hartmann 2015). Otherwise, rational analysis could simply be seen as the last breath of the autonomist dream of studying the mind independently from the brain.

In this paper, I assess the explanatory status of RA models by disentangling various explanatory contributions which have been attributed to them. By relying on the contrastive-counterfactual theory of explanation, I distinguish between three possible explanatory contributions such models could make: Uncovering (a) constitutive dependencies between parts and wholes, (b) environment-behavior dependencies, and (c) environment–optimal behavior dependencies. I treat the third alternative as the most promising source of new understanding provided by RA models. I argue that (c) should be interpreted as being explanatory not of human behavior as such, but of environmental affordances. Well conducted modeling of environmental affordances can complement mechanistic theorizing by providing means for understanding the possible space of behavior of agents.

2. Probabilistic cognitive modeling and rational analysis

2.1 Procedure of rational analysis

The idea of rational analysis modeling dates back to John Anderson's work on human memory and categorization in *The Adaptive Character of Thought* (1990). Having already worked on his ACT* cognitive architecture, the new methodology put forward in the book reflected Anderson's increasing worries that the research methods of the time could not really uncover cognitive mechanisms. Lacking a clear picture of what it is that cognitive mechanisms do (i.e. what the psychological explananda are), the available evidence of neural and algebraic level structures was insufficient to uncover the mechanistic architecture of the human mind (Anderson 1990, pp.23–26). Compared to bottom-up research strategies, rational analysis begins from the other end:

> *[…] We can understand a lot about human cognition without considering in detail what is inside the human head. Rather, we can look in detail at what is outside the human head and try to determine what would be optimal behavior given the structure of the environment and the goals of the human. (Anderson 1990, p.3)*

According to Anderson, careful mathematical modeling of the environment/task structure combined with an assumption about the optimality of human behavior leads to a new self-standing research strategy for understanding the mind: *"As this book is evidence, a rational analysis can stand on its own without any architectural theory"* (ibid.). By providing a precise model of what the mind does as a well-adapted system, rational analysis can constrain the search space for cognitive mechanisms, and put the scientific study of the mind on a firm foundation.

This view of the role of computational modeling immediately brings to mind Marr's (1982) account of multi-level theorizing in the mind sciences. However, whereas Marr provides no systematic model for building computational-level theories, RA modeling has predominantly proceeded according to the six-step modeling cycle proposed by Anderson (1990. p.29):

1. Specify precisely the goals of the cognitive system
2. Develop a formal model of the environment to which the system is adapted
3. Make minimal assumptions about computational limitations
4. Derive the optimal behavior function, given items 1 through 3
5. Examine the empirical evidence to see whether the predictions of the behavior function are confirmed
6. Repeat, iteratively refining the theory

These steps embody an account of how a large part of probabilistic cognitive modeling is done. However, two further assumptions should be made explicit. First, the derivation of optimal behavior in steps 2-4 typically employs *probability calculus* (not logic) as the normative baseline theory of rational behavior. Secondly, the connection between model predictions (step 4) and observed behavior of humans (step 5) is mediated by an assumption about the *optimality of the observed behavior* (see quoted passage above).

Below I illustrate this process with an example. However, a comment on the status of the approach in cognitive science is in place: Not all probabilistic modelers endorse the rational analysis framework (cf. Danks 2015; Sakamoto et al. 2008; Brighton & Gigerenzer 2008). Focusing on RA is useful for two reasons, however. Rational analysis is undeniably influential, and its core commitments have been endorsed a large group of well-known modelers (e.g., Anderson 1990; Oaksford & Chater 1994, 2007; Griffiths & Tenenbaum 2009). A further advantage of focusing on RA has to do with the fact that often the theoretical commitments of mathematical modelers are hard to pin down. In some cases, this is surely due to the modelers themselves not being clear of where their commitments (about explanatoriness, optimality, etc.) lie. Rational analysis provides a clear account of the conceptual foundations of probabilistic cognitive modeling, and therefore the following discussion is potentially helpful for challenging the methodological quietism among probabilistic modelers.

2.2 Oaksford and Chater on the Wason selection task

To illustrate the rational analysis process, I now briefly introduce Mike Oaksford and Nick Chater's (1994, 2007) analysis of the Wason selection task. Being a relatively simple model, it is a good device for illustrating the conceptual basis of RA modeling.

Wason selection task is one of the most famous laboratory experiments discussed in the literature on human rationality. In the original form of the task, subjects are given four cards, each of which has a letter on one side and a number on the other. The subjects' task is to determine whether the rule *"If there is a vowel on one side of the card (p), then there is an even number on the other side (q)"* holds. More precisely, subjects are asked to select all those cards, but only those cards, which would have to be turned over in order to discover whether the rule is true for the combination of cards they were given. The famous finding from the task and its several replications is that only a small minority of the subjects (less than 10%) select the correct cards (vowel, odd number) corresponding to the falsifying instance. Judged in the light of logic, most subjects fail to perform in a rational way.

Oaksford and Chater (O&C) challenge the irrationality claim by arguing that logic-based theories of inference and rationality misrepresent people's behavior in the task. O&C's own *information-gain model* of the situation argues that the apparently irrational behavior can be understood as the optimal way of decreasing uncertainty regarding the hypotheses studied. The gist of O&C's reinterpretation of the selection task is that instead of engaging in deductive reasoning, subjects interpret the task as inductive one. They do not try to falsify the rule, but instead they try to determine which of two hypotheses holds:

(a) Independence hypothesis **H_i:** $P(q \mid h) = P(q)$ or

(b) Dependence hypothesis **H_d:** $P(q \mid p)$ is high, higher than $P(q)$.

Being initially equally uncertain about both hypotheses, subjects aim to reduce this uncertainty as much as possible by turning as few cards as possible.

The rational analysis proposed by O&C relies on three basic starting points:

(1) Higher cognition can be modeled as probabilistic (Bayesian) computation

(2) The likelihoods and prior probabilities required by the model can be acquired from the analysis of the environment structure

(3) Behavior of human agents constitutes an optimal response to the task.

The Bayesian model of the situation is constructed roughly as follows.[1] To formalize the idea of uncertainty reduction, O&C adopt the optimal data selection paradigm, and interpret uncertainty reduction as optimization of *expected information gain*. Expected information gain $E[I_g]$ from turning over a card is defined as $E[I(H_i|D) - I(H_i)]$.[2] The Shannon information terms $I(H)$, in turn, are a function of the probabilities of the hypotheses before and after observing data, $P(H_i)$ and $P(H_i|D)$. These required posterior probabilities can be calculated from the likelihoods $P(D|H)$ and the priors by applying the Bayes rule. As the initial priors were set to be equal (.5), the rest of the crucial model specification is built into the likelihood functions, which describe the nature of the four-card task. Oaksford and Chater (1994, Table 1) show in detail how the required likelihoods can be read off the contingency tables describing the two hypotheses.

From these derivations, it follows that the crucial parameter values determining the optimality of behavior are the base rates of p and q. These probabilities describe how often positive instances of

[1] For mathematical details, see Oaksford & Chater 1994, 2007.

[2] Uncertainty (Shannon information) $I(H_i)$ given n mutually exclusive and exhaustive hypotheses (H_i), is $-\sum_{i=1}^{n} P(H_i) \log_2 P(H_i)$.

the antecedent and consequent of the rule appear in the environment. The expected information gain from turning the four cards depends on P(p) and P(q) in the following way:

- P(q) is small → P card is informative
- P(p) is large → Not-q card is informative
- P(p) and P(q) are small → Q card is informative
- Not-p card is not informative

How should these base rates, then, be determined? Instead of attempting to somehow measure the base rates of vowels and consonants in a relevant environment, O&C cite various intuitively plausible justifications for their *rarity assumption*. Relying on the observation that categories in language cut the world quite finely, the rarity assumption states that, generally, P(p) and P(q) are low in most situations.[3] Under rarity, O&C conclude, the q card is more informative than the not-q card. Hence, the model concludes that highest expected information gain is achieved by turning p and q cards, exactly as a majority of the participants do. Actually, with the parameter values chosen by O&C, there's a very good fit between meta-analysis results about people's behavior in the standard form of the selection task, and the predictions of the model. Hence, by changing the normative model of rational behavior, O&C were able to explain away irrationality, and to show that experimental subjects' behavior is actually very close to optimal.

The model has received critical attention in the literature (cf. Oaksford & Chater 2009), but it serves our current purposes well. The model specification and the modeling assumptions are conceptually

[3] See Oaksford & Chater 1994, 2007, and 2009 for alternative justifications of rarity.

on a par with those in more complex Bayesian models. The complexity in such models often pertains to the structure and generation of hypothesis spaces, and the models often rely on computational tools (such as MCMC approximation methods) to make the calculations tractable. However, these mathematical complexities have no influence on the fundamental conceptual structure of the model. What is common to all such models is that the none of the components (hypothesis space, likelihood function, and priors) are interpreted in a psychologically realistic way as mental representations (Jones & Love 2011). Instead, they stand directly for properties of the environment. Furthermore, data about human behavior is not fed into the model specification to empirically calibrate the model. Instead, it is only used to test model predictions. Hence, in this sense, the rational analysis of the selection task is an illuminating example of the theoretical and conceptual assumptions of computational probabilistic modeling.

3. What rational analysis models fail to explain

A shared starting point for many accounts of scientific explanation has been to distinguish explanation from other epistemic activities (e.g., description and prediction) by pointing out that explanations offer information of a specific kind. Explanations show *how* or *why* something happened or obtains. According to a now widely accepted approach, the knowledge that allows one to answer such questions concerns change-relating counterfactual dependencies between the relata in the explanation.

Stated generally, according to this contrastive-counterfactual theory of explanation, explanatory information has the following form (Woodward 2013; Ylikoski & Kuorikoski 2010):

{CC} *x [x'] because of y [y']* (variable X takes the value x instead of x' because Y has the value y instead of y')

In this account, being able to explain can be captured by being able to correctly answer what-if-things-were-different questions, i.e. questions of how changes in explanans variables lead to changes in the explanandum variable. In addition to being a sufficiently general account of explanation, the contrastive-counterfactual theory suits the purposes of this article well, because it does not necessarily tie the notion of explanation to that of causation. That is, although the 'because' in {CC} is typically understood as referring to causal dependency, the account does not rule out the possibility of non-causal explanation (Woodward 2013; Pincock 2015; Rice 2015): If there are ways of defining the notion of invariant dependency in non-causal situations (e.g. for mathematical dependencies), the contrastive-counterfactual theory could be applied to non-causal explanations as well. Hence, the theory of explanation casts the net wide enough to give RA models a fair chance of being explanatory.

A further advantage of treating explanations as answers to questions is that it allows us to make more precise the possible explanatory claims made by RA modelers. I suggest that there are at least three different kinds of objective dependencies that RA models could be said to track: (1) constitutive dependencies between parts and wholes, (2) environment-behavior dependencies, and (3) environment–optimal-behavior dependencies. In the rest of this section, I argue that in most cases of RA modeling, there are good reasons to conclude that the models do not have genuine explanatory import with respect to the two first kinds of dependencies.

3.1 Constitutive what-ifs

The notion of mechanism has acquired a central position in the philosophical debates concerning explanation in the life sciences. A clear expression of the mechanistic viewpoint has recently been given in the *model-to-mechanisms mapping (3M) requirement* by Kaplan and Craver (2011). According to the requirement, dynamical and mathematical models in systems- and cognitive

neuroscience explain a phenomenon only if there is a mapping between elements in the model and elements in the mechanism for the phenomenon. As the example discussed above suggests, rational analysis models provide no such mapping. They are agnostic about algorithmic and implementation level details, and intentionally so. Does this mean they cannot be explanatory?

First, as Kaplan and Craver themselves admit, their argument ultimately relies on shared norms about explanatoriness in the neuroscience community, and their account of explanation as construction of multi-level mechanisms reflects these norms. However, if such norms do not hold among probabilistic cognitive modelers, it is not obvious why they should abide by the 3M requirement.

Instead, if we understand explanation according to the contrastive-counterfactual theory, Kaplan and Craver's argument seems less disastrous: RA models obviously do not provide information about constitutive and causal dependencies in multi-level mechanisms, but according to the account, this does not rule out the possibility of RA models tracking some other kinds of objective dependencies, e.g. those holding between relata described in computational-level terms. Furthermore, a proponent of RA need not (and should not) claim that adding mechanistic detail never improves a computational explanation. To defend explanatoriness of RA models, a far weaker claim suffices, one stating that there can be explanatory contributions which do not rely on information from uncovering causal mechanisms.

3.2 Environment–behavior what-ifs

A second kind of explanatory question answered by an RA model could be "how would the behavior of the cognizer change when the cognitive task changes in some particular way?" That is, a RA model could uncover objective dependencies between properties of the environment and the

behavior of cognizers. For example, O&C's model can be used to derive predictions of what the behavior of the subjects in the Wason tasks would be, were P(p) and P(q) to take a range of values.

It is here that the optimality assumption of RA becomes crucial. To predict how human behavior would change in response to changes in the task, without knowing anything about the algorithms and processes which produce behavior, RA relies on the assumption that humans are well-adapted to their environments: If we assume that human behavior is optimal (or approximates optimal behavior) across a large variety of environments, the predictions derived from the RA model (step 4 of the analysis procedure) should in fact apply to that behavior.

Given that human (ir)rationality has been the topic of a longstanding debate in philosophy and psychology, it is not surprising that the optimality assumption has drawn a lot of criticism (cf. Jones & Love 2011). Although proponents of RA are correct in arguing that some degree of rationality of target behavior is required for us to even perceive it as intentional action, the modest levels of rationality needed hardly license the strong optimality assumptions in RA models. Neither do evolutionary arguments provide support for strong optimality claims: Although natural selection is a source of design and adaptedness, evolution is not guaranteed to produce globally optimal solutions – merely a local comparative advantage is sufficient for evolutionary solutions to survive.

Being aware of these problems, proponents of RA have avoided appealing to evolutionary defenses of the optimality assumption. Instead, they justify optimality by relying on an analogy to behavioral ecology and economics, where similar assumptions are commonly made (Chater et al. 2003). I believe, however, that the analogy breaks down due to a crucial dissimilarity between these fields: Both in biology and economics, rationality claims typically concern aggregate behavior, not that of

individual agents. Due to the disanalogy, I do not see how appealing to economics or biology could be a viable way to justify optimality assumptions in RA modeling.

These problems with general defenses of the optimality assumption suggest that perhaps optimality should be examined more locally. What kind of evidence should be obtained to justify the optimality claim in the case of a particular cognitive task? It seems that to support an objective dependency between environment and behavior, we should gather data about human behavior in a task *across a range of parameter values* describing various different environmental states. If human behavior fits the predictions made by the model across a range of conditions, that would appear to be rather strong evidence of optimality.[4]

Existing RA models rarely employ such cross-environmental data. First of all, many models not rely on any actual measurements of environment parameters (cf. Jones & Love 2011). Instead, they use plausible-sounding assumptions or analogies. For example, in O&C's selection task model, the base rates for p and q originated in such analogical reasoning. Similarly, Anderson's (1990, ch. 2) early model of memory relied on data about library borrowings to model usage of memory structures, and Griffiths et al. (2007) use Google PageRank to predict fluency of recall. Models devoid of good quality empirical data should be considered as toy models (at best), incapable of uncovering actual properties of cognitive environments.

[4] Note, however, that such empirical evidence for optimality would make the theory-based optimality assumption unnecessary.

Furthermore, as Marcus and Davis (2013, Table 1) observe, Bayesian modelers have been selective in the results that they report from experimental tasks. They only report ones where human behavior follows the model and ignore cases where its not optimal. Although some of the most recent models show some improvement in these respects, generally in RA models there is little evidence that could support knowledge of the needed invariant environment-behavior counterfactuals.

4. Rational analysis and the logic of the situation

Finally, let us think about the epistemic value of a RA model if we drop the optimality assumption. Assume that we have a rational analysis model with (i) well-specified task structure, (ii) parameter values based on empirical measurement of the environment, and (iii) an account of computational costs and limitations. What such a model could do is to link combinations of parameter values to best possible behavioral choices in those situations. Is this not a kind of objective what-if dependency? However, consider what the relata of such a dependency are. The model tells what the optimal behavior would be, given a particular combination of environmental conditions and computational limitations. Such counterfactuals do not say anything about actual human behavior. Instead, they increase our understanding of the environmental affordance, or, the logic of the situation (Popper 1963).

What mathematical models of affordances – the opportunities the environment offers for the agent – can help us understand is the possible space of behavior for cognitive agents. They show what a hypothetical rational agent would do in different situations. For what purposes could such information be useful? First, were we to design artificial cognitive systems with a particular cognitive task in mind, these systems should approximate the optimal behavior specified by the model. For example, in the selection task, *if* we are interested in reducing our uncertainty, O&C's

model tells us something non-trivial: It reveals the best choices of cards under different values of base rates for p and q.

Secondly, as in economics, rational models can act as normative baselines to which human behavior can be compared. As Sloman & Fehrbach (2008) argue, often it is just as interesting to find out that behavior does not conform to the norm than when it does. Finding out where and how systems malfunction is an efficient way to learn about them.

However, in neither of these uses are RA models employed to directly explain human behavior. Instead, they function as inferential aids which help to map the possible space of action for agents when faced with a particular task. Herein lies perhaps the hardest evidential problem faced by rational analysis. How do we know what the mind really does in some situation, i.e. where do the functional hypotheses in step 1 of RA come from? For example, how would O&C defend their probabilistic construal of the selection task against an adamant falsificationist? Available empirical evidence can hardly decide the issue: Where O&C see optimal behavior, the falsificationist sees well-known inferential blunders.[5] Marcus and Davis (2013) argue that similar problems of model selection plague several other RA models as well.

The difficulty seems to come down to the fact that the cognitive tasks and the affordances available for an organism depend on its "life space" – not the physically objective world in its totality, but reality filtered through the organism's needs, drives and perceptual apparatus (Simon 1956).

[5] What makes O&C's model selection seem even more ad hoc is that they do not explain different versions of the selection task (e.g., the deontic selection task) by using the same model, but instead they introduce modified versions for each of the variations.

Therefore, there is no reason to think that a mathematician's intuitions are a reliable guide to what the cognitive tasks of human agents are. Ad-hocness in model selection, in turn, raises serious worries about the *relevance of RA modeling*: Constructing detailed mathematical models of potential affordances is of little interest unless they can be shown to be ones humans actually track.

This leads me to my conciliatory conclusion. As suggested both by the connectionist rivals of RA and proponents of multi-level mechanistic explanation in philosophy (McClelland et al. 2010; Bechtel & Richardson 2010), functional hypotheses in cognitive science must be formulated in an iterative process between bottom-up and top-down research strategies. On the one hand, knowledge about perceptual and computational constraints of organisms mostly originates in bottom-up research on the mind-brain, and this knowledge should be allowed to constrain RA models. In this sense, Anderson's and O&C's claims about the self-standing explanatory role of RA are not vindicated by my analysis. However, the discussion on mechanistic explanation has been downward-looking in spirit, and modeling the environment within which cognitive mechanisms function has not received enough attention. Here RA models can complement mechanistic theories of cognition by providing precise mathematical models of the task and the environment. For example, as Chater et al. (2003) point out, a correctly formulated rational analysis can show why it is that some simple approximating heuristic is successful in solving a computationally complex task.

4. Conclusions

I have argued that given a sufficiently broad account of scientific explanation, there are several possible ways in which probabilistic modeling could increase our understanding of the mind. However, the strictly-computational methodology embodied in the six-step formula of rational

analysis has led to theorizing which often fails to reliably uncover genuine explanatory dependencies. The shortcomings of RA are evidential in nature: The nature of the data, and the way it is used in model construction allows too easy curve fitting, and it is insufficient for reliable counterfactual inference.

My new proposal about the epistemic role of RA models without the problematic optimality assumption is that they can be understood as models of environmental affordances. Interpreted in this way, RA models do not actually provide information about the mind works, or hypotheses about cognitive functions (Zednik & Jäkel 2014). Instead, they map the possible cognitive space of action for an organism. The explanatory contribution of such information is best worked out as constituting a part of a non-reductionist mechanistic research programme.

References

Anderson, John. 1990. *The Adaptive Character of Thought*. Hillsdale: Lawrence Erlbaum Associates.

Bechtel, William. and Richardson, Robert. 2010. *Discovering Complexity.* The MIT Press.

Brighton, Henry. & Gigerenzer, Gerd. 2008. Bayesian brains and cognitive mechanisms: harmony or dissonance? In Chater & Oaksford (eds.) *The Probabilistic Mind*. Oxford University Press.

Chater, Nick, et al. 2003. Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes,* 63–86.

Chater, Nick. & Oaksford, Mike. (eds.) (2007). *The Probabilistic Mind.* Oxford: Oxford University Press.

Colombo, Matteo. & Hartmann, Stephan. (2015). Bayesian cognitive science, unification, and explanation. *British Journal for the Philosophy of Science.*

Danks, David. 2015. *Unifying the Mind*. MIT Press.

Griffiths, Thomas., Steyvers, Mark., & Firl, Alana. 2007. Google and the mind. *Psychological Science*, 1069–1076.

Griffiths, Thomas. & Tenenbaum, Joshua. 2009. Theory-based causal induction. *Psychological Review*, 661–716

Jones, Matt. & Love, Bradley. 2011. Bayesian Fundamentalism or Enlightenment? *Behavioral and Brain Sciences*, 34, 169-231.

Kaplan, David. & Craver, Carl. 2011. The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, 78, 601-627.

Marcus, Gary, & Davis, Ernest. 2013. How robust are probabilistic models of higher-level cognition? *Psychological Science*, 24, 2351–2360.

Marr, David. 1982/2010 *Vision*. W.H. Freeman/MIT Press.

Oaksford, Mike, & Chater, Nick. 1994. A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608-631,

— 2007. *Bayesian Rationality: The probabilistic approach to human reasoning*. Oxford University Press.

— 2009. Precis of Bayesian Rationality. *Behavioral and Brain Sciences*, 69–120.

Pincock, Christopher. 2015. Abstract explanations in science. *British Journal for the Philosophy of Science* 66, 857-882.

Popper, Karl. 1963. Models, instruments, and truth. Manuscript. Karl Popper Collection at the Hoover Institution Archives at Stanford University.

Rice, Collin. 2015. Moving beyond causes: Optimality models and scientific explanation. *Noûs* 49, 589-615.

Sakamoto, Yasuaki., Jones, Matt. & Love, Bradley. 2008. Putting the psychology back into psychological models**.** *Memory & Cognition*, 36, 1057–1065.

Simon, Herbert. 1956. Rational choice and the structure of the environment. *Psychological Review*, 129–138.

Sloman, Steven, & Fehrbach, Philip. 2008 The value of rational analysis: as assessment of causal reasoning and learning. In *The Probabilistic Mind.*

Woodward, James. 2013. Mechanistic explanation: Its scope and limits. *Proceedings of the Aristotelian Society Supplementary Volume,* lxxxvii: 39–65.

Ylikoski, Petri., & Kuorikoski, Jaakko. 2010. Dissecting explanatory power. *Philosophical Studies*, 148, 201–219.

Zednik, Carlos. & Jäkel, Frank. 2014. How does Bayesian reverse-engineering work? In P. Bello et al. (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 666-671).

Inherent Complexity: a problem for Statistical Model Evaluation

Jan-Willem Romeijn
University of Groningen

Abstract

This paper investigates a problem for statistical model evaluation, in particular for curve fitting: by employing a different family of curves we can fit a scatter plot almost perfectly at apparently minor costs in terms of model complexity. The problem is resolved by an appeal to prior probabilities. This leads to some general lessons about how to approach model evaluation.

1 Introduction

Theories often interface with empirical fact through a statistical model, namely a collection of hypotheses that each determine a probability distribution over possible observations. Most statistical inference is carried out on the basis of a model, for example by getting the data to choose among the hypotheses in it, or by redistributing the probability assignment over the hypotheses in the model.

Curve-fitting is an instance of statistical inference. For example, the yearly number of car accidents with claimable damage follows a Poisson distribution, whose characteristics depend on the total distance covered by the vehicle. What determines the statistical model is the exact functional dependence of frequency

on distance. Since vehicles that do not cover any distance will not incur any damage, the intercept will be zero. One statistical model may be that the dependence is linear, so that the hypotheses in the model differ in the slope of the line that relates distance to expected number of accidents. Another statistical model might postulate a more complicated relation between distance and expected number of accidents, e.g., a quadratic dependence, perhaps with the idea that long-distance drivers have proportionally fewer accidents.

While models are typically chosen at the outset, sometimes they are under scrutiny themselves. For example, we might compare the linear and the quadratic models sketched above. Statistical model evaluation allows us to compare such models on a variety of performance measures. Model evaluation is important for scientists and philosophers of science alike. It allows scientists to submit their modeling assumptions to empirical testing, and thereby address the uncertainty over their theoretical starting points. And it gives philosophers of science a concrete and formally precise handle on a fundamental kind of uncertainty. Examples of model evaluation abound, ranging from climate science and ecology to psychiatry and computational archaeology. If philosophers can motivate and develop norms for dealing with model uncertainty, this will have direct implications for the practice of science.

This paper contributes to our understanding of the norms that drive statistical model evaluation. After an introduction into model evaluation tools in section 2, I present a new problem for them in section 3. I then offer a diagnosis of the problem in section 4. In section 5 I show that the problem can be avoided if we involve prior probability assignments in the model evaluation. Throughout I will mostly avoid mathematical detail, to leave more space for conceptual considerations.

2 Statistical model evaluation

The curve fitting problem sketched in the introduction may not seem statistical. Given a family of curves, we simply choose by minimizing the errors, i.e., the

sum of the discrepancies between curve and point. In the so-called least-squares approach, for example, the error is calculated as the sum of the squares of the vertical distance between point and curve. No model seems to be involved in this.

Underneath such a minimization procedure, however, we do find a statistical inference. One central modeling assumption is that the number of accidents N follows a Poisson distribution. A further assumption is that the mean of this distribution depends on the distance D covered by the vehicle,

$$P_\theta(\langle D, N\rangle) = \frac{(\lambda(D))^N}{N!}e^{-\lambda(D)}, \tag{1}$$

with $\lambda(D) = \theta_1 D + \theta_2 D^2$. Then we choose $\theta_1 > 0 > \theta_2$ for the quadratic model and $\theta_1 > 0 = \theta_2$ for the linear one. Note that the model dictates a distribution over N for all values of D but that it does not determine a probability distribution over the values of D itself. The distance D is an explanatory variable, and we presume that it is randomly sampled from a uniform distribution.

The data consist of m pairs of distances and numbers of accidents, collected in a scatter plot:

$$S_{DN} = \{\langle d_1, n_1\rangle, \langle d_2, n_2\rangle, \ldots, \langle d_m, n_m\rangle\}. \tag{2}$$

For any curve and associated hypothesis we can calculate the probability of a scatter plot, i.e., the likelihood of the hypothesis for the data, by multiplying the probability of all the points,

$$P_\theta(S_{dn}) = \prod_{i=1}^{m} P_\theta(\langle d_i, n_i\rangle). \tag{3}$$

A data point $\langle d, n\rangle$ lying outside the normal range for some hypothesis, e.g., with n high while d is low and θ_2 is too, will be improbable, and hence it will strongly decrease the likelihood of the hypothesis. To fit the curve we look for the value of θ, denoted $\hat{\theta}$, that makes the probability of the scatter plot maximal. Generally speaking, maximizing the likelihood of the curve will correspond to minimizing the distance of points to the curve under some notion of distance. Figure 1 offers an impression of what these curves may look like.

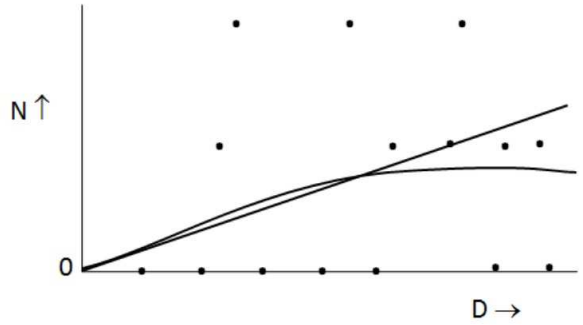


Figure 1: The polynomial curves fitted to the scatter plot.

Against this background it will be clear that evaluating the general shape of the curves, comparing linear and quadratic ones, is indeed part of statistical model evaluation. Note that I use the term "model evaluation", not the more often used "model selection". The selection of a model is a decision, and so involves decision-theoretic as well as inferential aspects. But in what follows I will only consider norms for the comparison of models from an epistemic standpoint.

A very common idea about model evaluation is that, next to the fit with data, it involves the complexity of the model. If a neat fit with the data is achieved by adding many bells and whistles, we are rightly reluctant to put our trust in it. We then say that the model is fitting to noise, or overfitting. In the example, the best fitting curve from the quadratic model will have a higher likelihood than the best curve from the linear model. But this is not to say that the quadratic curve is better. The question is whether the gain in fit weighs up against the cost of a more complicated model.

The extant model evaluation tools, most notably the various information criteria (Claeskens and Hjort, 2008), provide specific formats for this trade-off between simplicity and fit. The two most prominent tools, the Akaike and Bayesian information criteria or AIC and BIC for short, express the simplicity by means of the number of free parameters in the model (cf. Akaike, 1973; Burnham and Anderson, 2002; Raftery, 1995; Schwarz, 1978). The linear model

of the example has one free parameter, and the quadratic model has two. The ICs then differ in how they factor the number of parameters into the trade-off:

$$\mathrm{AIC}(\mathcal{M}_\theta) = 2\log\left(P_{\hat{\theta}}(S)\right) - 2\dim(\mathcal{M}_\theta), \quad (4)$$

$$\mathrm{BIC}(\mathcal{M}_\theta) = 2\log\left(P_{\hat{\theta}}(S)\right) - \log(m)\dim(\mathcal{M}_\theta), \quad (5)$$

in which $\mathcal{M}_\theta$ is the model parameterized by the vector θ, the number of free parameters is given by $\dim(\mathcal{M}_\theta)$, and $\hat{\theta}$ is the hypothesis in the model with maximum likelihood for the data D, so that $P_{\hat{\theta}}(S)$ is the likelihood of the maximum likelihood estimator for the data S. In the BIC the penalty for complexity is scaled according to the sample size of the data m.

The involvement of the complexity of models in their evaluation may seem intuitive on pragmatic or metaphysical grounds. A simpler model is easier to use, or we might think that the world itself is a simple place, perhaps because the Demiurge is an efficient or lazy being. The actual reason for the appearance of the complexity penalty in the ICs is epistemic though. Moreover, the motivation is different for the various information criteria on offer. For example, the AIC factors in the number of parameters as a result of approximating the expected Kullback-Leibler divergence to the true hypothesis. And for the BIC the penalty for complexity drops out of an approximation of the past predictive performance of the model, as measured by the marginal likelihood.

The number of model parameters surfaces repeatedly as a criterion for model evaluation, under a variety of epistemic good-making features of models. Very roughly, the underlying reason is that the predictions and general empirical claims of more complex models will be less robust and reliable. In a more complex model the same number of data points will be used to determine a larger number of parameter values, and so the available information will have to be spread more thinly. For the AIC this shows up in the stumpness of the likelihood function over the model, and in the BIC it appears as the stumpness of the posterior probability distribution within the model. The general idea is that we can always introduce an additional parameter that improves the best

fit in the model, but that we might then lack the data to properly back up a stable value for this additional parameter.

However, this intuition does not cover everything that is salient about complexity in model evaluation. There is another epistemic good-making feature, strongly related to complexity and the number of parameters, that needs to be taken into account when we compare models. This further feature concerns something like model size. It can be expressed by means of the prior probability distribution within the models, as the following model evaluation problem will reveal.

3 Cheap and almost perfect fit

Consider again the example of the scatter plot and the polynomial model. But instead of using the polynomial curves, as detailed above, imagine fitting the data with a model based on trigonometric functions, or sine curves for short. We use the Poisson distributions of Equation (1) but instead of choosing $\lambda(D)$ to be polynomial we choose

$$\lambda(D) = \alpha_1 - \alpha_1 \cos(\alpha_2 D). \tag{6}$$

Figure 2 gives an impression of the fit that may be achieved by the so-called sine model. Importantly, all the points in the scatter plot are given close to maximal probability, because they all end up sitting arbitrarily close to the curve, and hence to the mean for the distribution at the given distance D.

The key observation is that we have achieved this remarkable fit at the expense of only two parameters, α_1 and α_2. It is known that we can obtain a perfect fit to m data points with a polynomial curve of degree $m - 1$. But fitting any number m of points with two parameters seems inexplicably efficient. Clearly, if we were to apply model evaluation criteria like AIC or BIC, or indeed any other method in which complexity is expressed by the number of parameters, the sine model wins out on the quadratic model, and most likely also on the linear model. What is going on?

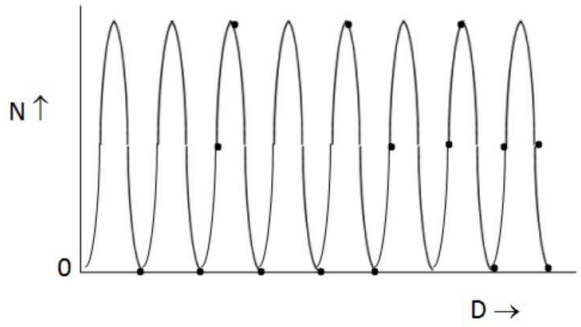


Figure 2: The sine curve that perfectly fits the scatter plot.

Before providing a diagnosis, let me emphasize that the claim that a near-perfect fit is always possible is mathematically non-trivial. In the remainder of this section I will provide more detail to substantiate it. Notably, the fit does not hinge on the assumption of any particular distribution, be it Poisson, normal, or otherwise, or on any particular format of the data, be it real numbers, integers or otherwise. Moreover, given that the scatter plot will manifest on a finite domain $0 < D < L$ we need not even suppose that the parameters are real valued: it is enough to consider sine curves with a period $^{L}/_{t}$ for $t \in \mathbb{N}$, as one does in a Fourier series. Despite all this, it turns out that there are always infinitely many almost perfect fits to a set of points. This abundance of solutions will turn out to be of crucial importance for the resolution of the problem.

Say that we have been given a scatter plot S_{dn} whose farthest points are at $d_i = L$ and $n_j = H$. For convenience we set $\alpha_1 = {}^{H}/_{2}$, but any α_1 larger than that will work too. Take any specific point $\langle d, n \rangle$ from the scatter plot, consider the curves with $\alpha_2 = {}^{L}/_{t}$ for increasing t, and ask: for what values of t does the sine curve intersect with the line $D = d$ in very close proximity to the value n? Observe that $d \in [{}^{kL}/_{t}, {}^{(k+1)L}/_{t}]$, and that the curve covers the whole of the range $[0, H]$ over this interval of D twice. If we allow for a discrepancy of ϵ between the curve and the value n and assume that d falls within the first

half of the interval, we must require that

$$\frac{L}{\pi t}\cos^{-1}\left(1-{}^{2n-\epsilon}/_{H}\right) \;<\; d-\frac{kL}{t} \;<\; \frac{L}{\pi t}\cos^{-1}\left(1-{}^{2n+\epsilon}/_{H}\right). \tag{7}$$

If d falls in the second half of the interval $[{}^{kL}/_{t}, {}^{(k+1)L}/_{t}]$, we require an analogous constraint. Because the slope of the cosine is bounded between -1 and 1, we may replace the above inequalities with

$$\frac{L}{\pi t}\cos^{-1}\left(1-{}^{2n}/_{H}\right)-\frac{\epsilon L}{\pi t H} \;<\; d-\frac{kL}{t} \;<\; \frac{L}{\pi t}\cos^{-1}\left(1-{}^{2n}/_{H}\right)+\frac{\epsilon L}{\pi t H}, \tag{8}$$

and similarly for d sitting in the second half of the interval. Consequently, for every t there is a specific region of length ${}^{4\epsilon L}/_{\pi t H}$ within the interval of length ${}^{L}/_{t}$ that includes d, for which the resulting error lies within an ϵ bound. The question merely is, for every separately t, whether d indeed lies within this specific region.

To establish when the latter obtains, we first recall that the d_i's from the scatter plot S_{dn} were randomly sampled from a uniform distribution over $[0, L]$. This means that the individual d from the sample is almost surely, i.e., with probability one, a random number. Consequently, there will be no pattern in how d shows up inside the intervals $[{}^{kL}/_{t}, {}^{(k+1)L}/_{t}]$ for increasing t. The locations of d are evenly distributed over all parts of this interval. Hence for any $\epsilon > 0$ there will be infinitely many t for which d will fall within the portion of length ${}^{4\epsilon L}/_{\pi t H}$ inside the interval of length ${}^{L}/_{t}$. The relative size of the region in which the curve is sufficiently close to the value n is constant for increasing t at ${}^{4\epsilon}/_{\pi H}$. And so there will be infinitely many sine curves that have an arbitrarily small error in fitting the point $\langle d, n\rangle$.

This suffices as an argument for there being an infinity of curves that fit any finite number of points almost perfectly. For a single point, the fraction of sine curves will tend to ${}^{4\epsilon}/_{\pi H}$. So for a set of m points that are randomly distributed over D, the fraction will tend to $({}^{4\epsilon}/_{\pi H})^{m}$. When making ϵ small and thus maximizing the likelihoods, the fraction of curves with good enough fit will be very small. But there will still be infinitely many fitting ones.

4 Diagnosis of the problem

The fact that there are infinitely many equally well-fitting sine curves incapacitates some of the standard model evaluation tools. The AIC, for example, is not defined for unidentified models. While being silent may be better than positively evaluating the intuitively incorrect sine model, a negative evaluation of the sine model seems preferable. Our discussion revolves around three observations: the sine model is not robust, counting parameters is a nontrivial affair, and the set of best fitting sine curves is not well-behaved. This sets us up for a solution to the problem along Bayesian lines in the next section.

First consider the robustness of the sine model. Imagine that we alter the scatter plot by slightly nudging a single data point. What will be the result if the curve is a polynomial of a given degree? Clearly, any curve that was fitted to the data will change a little as well. But the rough shape of the curve will not change a lot: a small change in data is matched by a similarly small change to the best fitting curve. By contrast, if the curve is a trigonometric function, then nudging a single data point slightly will radically alter the best fitting curve. It will lead to a completely new set of best fits. We might say that the sine model is too versatile, lacking robustness, or skittish: it is oversensitive to the smallest of changes in the scatter plot.

The AIC and BIC do not accommodate this feature of models. MDL-based model evaluation tools and extensions of the AIC and BIC fare slightly better. The Fisher information approximation (FIA) for example includes a so-called geometric complexity term, based on the Fisher information. One might say that this expresses model size in terms of how densely packed the model is with likelihood functions (Grunwald, 2007; Myung et al, 2000; Ly et al, 20XX). The term penalizes skittish models because they will in general cover a larger set of probable data patters: small changes lead to wildly different functions, and in this sense the skittish models are indeed packed densely. Furthermore, developing the AIC and the BIC, as discussed in Bozdogan (1987) and further references therein, we also encounter the Fisher information. So there are nat-

ural extensions of the AIC and the BIC that accommodate something of the skittishness.

However, in all of these refined methods, the contribution of the Fisher information (FI) term is not of the required order of magnitude to resolve the problem of the sine curves. Apart from the original AIC, the FI term is trumped by the term that captures complexity as the number of free parameters, and which grows with $\log(m)$. And the FI term cannot compete with the fit term, which grows with m in all model evaluation tools. For larger data sets the influence of the FI term on the model score therefore dwindles, so that the sines seem preferable after all.

A second observation concerns the deceptively low dimensions of the sine model: it seems to harbor an inherent complexity that is not expressed in the number of parameters. The sine model illustrates that model dimension is a fleeting notion. As a quick illustration, note that statistical parameters are often real numbers. But real numbers are such that we can package any amount of information into them. For example, a sufficiently complicated function will allow us to compact two real numbers in a single one, by constructing the numerical expansion of a number from two such expansions, e.g., $0.135\ldots$ and $0.246\ldots$ yield $0.123456\ldots$, and so on. While this sort of function is of course hopelessly contrived, it illustrates that counting statistical parameters does not give us a fair indication of model dimensions.

This general observation has been made about model evaluation criteria more often, for example in Bozdogan (1987), who proposes to adapt the AIC by involving the sample size, thereby bringing it closer to the BIC. His motivation for adapting the penalty term is, by and large, that the notion of complexity is not adequately captured by the dimension term in the original AIC. Similar sentiments are expressed in Balasubramanian (2005) who develops minimum description length (MDL), and in Romeijn and van de Schoot (2008); Romeijn et al. (2012) who investigates and extends the BIC. The latter two point to a more general notion of model size as a component of complexity. However,

while these proposals are in the right direction, the adapted versions of AIC, BIC, and MDL still give the number of parameters a central role.

A more promising method for dealing with the problem of the sine curves is offered by the so-called Deviance information criterion, or DIC for short (Spiegelhalter et al, 2002). The DIC was originally designed for comparing hierarchical Bayesian models, in which the number of free parameters is not clearly defined. Central to the DIC is the so-called deviance, i.e., the reduction in surprise due to estimation, which can be thought of as a degree of overfitting. The penalty for complexity in the DIC is given by the effective number of parameters, which is based on the notion of deviance. However, in this paper I will not investigate in detail how the DIC responds to the sine model.

A final observation brings us closest to the ultimate reason that trigonometric curves are problematic for the purpose at hand. Note that both polynomial and trigonometric curves can be used as a basis for the space of functions on a finite domain, in the algebraic sense that they parameterize that space: we can write down functions by their Taylor or Fourier series. We can collect the curves that almost perfectly fit some scatter plot into a set within the space of functions. But for the Taylor and Fourier series this set will look very different. In the Taylor parameterization, the set is a well-behaved region sitting somewhere in the linear combination of at least m axes. But the set of well-fitting curves will look much wilder and disjointed in the Fourier parameterization, disjointed and intersecting with distinct axes rather than being lumped together.

The implications of this are best brought out through a variant of the robustness discussed above, namely by considering what happens if we add a point in the scatter plot. The original polynomial curve will not change too radically: the region of well fitting curves shifts slightly. By contrast, the set of best fitting sine curves alters significantly with the addition of a point, not so much by being relocated but rather by being constrained severely. There are infinitely many sine curves that fit the scatter plot, but almost all of those curves will miss the additional point by a stretch, and so be eliminated from the set of well

fitting curves. The solution of the problem hinges on exactly this elimination of hypotheses.

5 Priors to the rescue

This section develops a particular response to the problem of the sine curves. It relies on so-called Bayesian model selection, or Bayesian model evaluation (BME). Following BME the sine model looses against polynomial models because of the specific failure of robustness introduced above.

The message of this section is not that we should embrace BME as the new standard in model selection. I will not make a systematic comparison with other model evaluation criteria and their relation to the salient notion of robustness. Looking at the solution that BME provides and the central role for the so-called marginal likelihood in BME, we might expect that other approaches in which the marginal likelihood is central, e.g., the DIC and MDL-based criteria, will also provide a solution. Because we can only compute something like a marginal likelihood if we adopt some version of a prior within the model, the central point of this section is rather that solutions will have to rely on priors of some kind.

The central idea of BME is to compare models by their posterior probability assignment:

$$\frac{P(\mathcal{M}_1|S)}{P(\mathcal{M}_2|S)} = \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)} \times \frac{P(S|\mathcal{M}_1)}{P(S|\mathcal{M}_2)}. \tag{9}$$

Assuming an equal prior for the models $\mathcal{M}_i$, the posteriors are completely determined by the ratio of the so-called marginal likelihoods,

$$P(S|\mathcal{M}_i) = \int_\Theta P(H_\theta|\mathcal{M}_i) P(S|H_\theta \cap \mathcal{M}_1)\, d\theta, \tag{10}$$

in which Θ is the parameter space. The likelihoods $P(S|H_\theta \cap \mathcal{M}_1)$ are a different notation for the $P_\theta(S)$ of the foregoing. Notice that the prior within the model, $P(H_\theta|\mathcal{M}_i)$, plays a key role in the computation of the marginal likelihood. Many approaches to model evaluation rely on the marginal likelihood of the model, including the BIC, the DIC, and MDL-based approaches. All these approaches must use some notion of a prior.

Now recall how sine curves manage to fit any scatter plots almost perfectly, in particular that there are infinitely many such curves. With the addition of a new point this set of best fitting curves will loose a large number of members, and this will severely impact the marginal likelihood of the sine model. Following Equation (3), we see that the likelihood of sine curves that retain their fit be multiplied by a maximal probability for every new data point. But this only holds for a small fraction $4\epsilon/\pi H$ of the sine curves. The fraction $1 - 4\epsilon/\pi H$ of sine curves will be multiplied by a factor that falls far short of the maximum probability.

By comparison, the likelihoods of the well fitting polynomial curves will pick up a factor that is somewhat lower than the maximal probability for each point, though not falling very far short of the maximum. Importantly, this high but not maximum factor will apply to a set of curves that is more or less stable and that will accumulate more and more probability with the addition of data points. Consequently, the overall factor picked up by the marginal likelihood of the polynomial models will tend to this high but not maximum factor.

To put this in a more mathematical format, say that the average factor picked up by the likelihood of a sine curve outside of the set of best fitting curves is U, that the same factor applies to badly fitting polynomial curves, that the factor for a well fitting polynomial is V, and for a best fitting sine curve W, so that $U < V < W$. For the sine curves we obtain

$$P(\langle d_{m+1}, n_{m+1}\rangle | \mathcal{M}_{\text{Sine}} \cap S_{dn}) \approx \left(1 - \frac{4\epsilon}{\pi H}\right) U + \frac{4\epsilon}{\pi H} W, \tag{11}$$

which is arbitrarily close to U. For the polynomial curves we will have

$$P(\langle d_{m+1}, n_{m+1}\rangle | \mathcal{M}_{\text{Poly}} \cap S_{dn}) \approx (1 - R_m)U + R_m V, \tag{12}$$

in which R_m tends to 1 for increasing m so that the factor tends to V. The result is that the sine model performs less well than the polynomial model on the BME criterion. On BME, therefore, the inherent complexity of the sine curves is adequately factored in.

It will be insightful to return to the observations that the set of well fitting sine curves is skittish. Well fitting polynomial curves of a given degree are concentrated in a particular region within the model, in which posterior probability will accumulate when data size increases: all of them will respond to new data points in roughly the same way. By contrast, with every new data point a small fraction of the well fitting sine curves is multiplied by a high likelihood, while a large portion picks up a low factor. It expresses the skittishness of sine curves that such a large portion of curves is suddenly far off in their prediction.

We can also convert this reasoning to arrive at the observation about model size. Judged from the prior probability distribution within the sine model, the set of well fitting curves is very small indeed: after m points it has decreased to $({}^{4\epsilon}/_{\pi H})^m$. But considering the prior within the polynomial model, the set of well fitting curves retains a reasonable size. What this signals is that the sine model, although it has only two free parameters, has many more different statistical hypotheses packed into it. It is versatile at the cost of a particular kind of robustness. The use of a prior within the model enables us to bring this kind of robustness out.

6 Conclusion

We cannot turn the foregoing into an argument for BME: other model evaluation criteria may also have a response to the problem at stake. But there are several general lessons to take away. One is that we must never mistake the number of parameters in a model for its actual complexity. A related lesson is that we must not forget the deeper motivations for the model evaluation tools, i.e., the good-making features that the tools are based on. Concentrating on those features will guide us to a better understanding of our evaluations.

Another general general lesson ties in with earlier work on the role of size in model evaluation (Romeijn et al., 2012), and indeed with scientific methodology as a whole. In the solution of the problem with the sine model, we can recognize a Popperian theme. Models that allow for fewer possible data patterns are

preferable to those that allow for a very wide range of data patterns. To express some notion of model size in our evaluations, we have to adopt some measure over the space of distributions over data. So we must involve something akin to a prior.

There is, however, a problem with the idea that we can objectively determine how densely distributions are packed together in a model. To say that a set of distributions shows a wide variety in the data patterns that it can adapt to, we need to presuppose a notion of similarity among data patters or, more generally speaking, a metric over sample space. In this paper that metric was adopted implicitly, as part of the way in which we depict and conceptualize the data. This dependence on the metric of the sample space points to a potential subjectivity in adjudicating between statistical models, or at least a reliance on a natural conceptualization of the sample space. This idea deserves to be studied in its own right.

Acknowledgements

The author wishes to thank Elliott Sober and Tom Sterkenburg, as well as audiences in Santiago de Compostella, Padova, Helsinki, Gent, and Groningen.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *2nd International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Akademiai Kiado, Budapest, pp. 267–281.

Balasubramanian, V. (2005). MDL, Bayesian inference, and the geometry of the space of probability distributions. In *Advances in Minimum Description Length: Theory and Applications*, P. J. Grunwald et al. (eds.), pp. 81–99. MIT Press, Boston.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion. Psychometrika 52(3), 345–370.

Burnham, K.P. and D.R. Anderson (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. New York: Springer.

Claeskens, G. and N. L. Hjort (2008). Model selection and model averaging, Cambridge: Cambridge University Press.

Grunwald, P. (2007). The Minimum Description Length Principle. MIT press, Cambridge (MA).

Ly, A. J. Verhagen, R. Grasman, and E-J. Wagenmakers (20XX). A Tutorial on Fisher Information, unpublished manuscript.

Myung, J. et al. (2000). Counting probability distributions: Differential geometry and model selection. Proceedings of the National Academy of Sciences 97(21), pp. 11170–11175.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, pp. 111–163.

Romeijn, J. W. and R. van de Schoot (2008). A Philosophical Analysis of Bayesian model selection. In Hoijtink, H., Klugkist, I., and Boelen, P. A. (2008). *Bayesian Evaluation of Informative Hypotheses*, Springer, New York.

Romeijn, J.W., R. van de Schoot, and H. Hoijtink (2012). One size does not fit all: derivation of a prior-adapted BIC. In Dieks, D., W. Gonzales, S. Hartmann, F. Stadler, T. Uebel, and M. Weber (eds.), Probabilities, Laws, and Structures. Berlin: Springer.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6, pp. 461–464.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society B*, 64, pp. 583–639.

Introduction

The scientific realist claims that the physical sciences provide, or aim to provide, a true description of the underlying reality behind the manifest world of experience.[1] But much of physical theory is expressed in the language of mathematics, and if any form of scientific realism is to be grounded in a mathematical science, it is essential that the realist provide an account of how mathematics is applied to the physical world. Philosophers of applied mathematics often explicate the use of mathematics in the physical sciences in terms of the concept of a representation. In the popular "mapping account" of applied mathematics, it is argued that we use mathematics to represent certain physical structures (Brown, 1999 and 2012; Pincock, 2007 and 2012). The basic idea is that we identify a physical structure in the world and then map it onto the appropriate mathematical structure within our scientific theories (Brown, 2012; 6-7, and Pincock, 2012; 27-29). Formally, a representation occurs when a morphism can be specified between a relational system in the physical world and a mathematical structure. Based on the popularity of the mapping account of applied mathematics, it worth taking the time to see if this approach can provide a viable foundation for scientific realism. At first glance, scientific realism and the mapping account appear to be a match made in heaven. Following the mapping account, the realist would be able to suggest that mathematics is successfully applied when the relations that hold within a physical system are correlated with the appropriate mathematical structure.[2] However, the mapping account of applied mathematics has met with wide-ranging criticism (e.g. van Fraassen, 2008; Batterman, 2008; Bueno and Colyvan, 2011; and Berkovitz, 2015). For the realist, the most pressing concerns with the mapping account pertain to how a physical structure is identified and represented as a mathematical structure.

The mapping account is appealing to the scientific realist specifically because it is a variant of the copy theory of representation. In the copy theory of representation, we rep-

[1] Alternatively, this sentence could be made compatible with the usual concessions to approximate truth.

[2] Pincock notes that this is the condition for the successful application of mathematics within the mapping account (Pincock, 2012; 28).

resent an object (or physical relation) by copying it, or an aspect of it, onto the intended representation. Such an account would allow the scientific realist to refer to the mathematical structure of a scientific theory as a copy of the physical structure in the world. However, any copy theory of representation is subject to Goodman's (1976) criticism. Goodman notes that the copy theory of representation is "stopped at the start" by an inability to identify exactly what is being copied by the representation relation (Goodman, 1976; 9). In Goodman's view, we do not copy the object or relation itself, but rather how the object or relation is conceived. When we conceive of an object or relation, we construe or interpret it and "[i]n representing an object [or relation], we do not copy such a construal or interpretation–we achieve it" (Goodman, 1976: 9).[3] The problem is that the world does not come 'carved at its joints'.[4] Rather, the joints are constructs of our conceptual systems, i.e. theories.[5] Although Goodman's general philosophical position is controversial, his point is clear in the case of theoretical physics, where the "physical structure" that is being represented is not readily apparent. In fact, the "physical structure" itself has to be constructed out of a mathematical theory of the world. The constitutive role that mathematics plays in the physical sciences presents a serious problem for the mapping account of applied mathematics. If mathematics is applied in the construction of our physical conception of the world, then it has certainly overstepped the boundaries of the copy theory of representation. Rather, representation becomes essential to the very construction of the physical structure that is at the foundation of our scientific theories.

The same issue can be viewed from another perspective. At the heart of the mapping account lies a relation that maps a physical structure onto a mathematical structure. This relation is defined as a morphism, which is a mathematical relation.[6] The problem is that a

[3]The insertion of the phrase "or relation" is supported by Goodman's footnote on page 5.

[4]This issue has been recently addressed in the context of the mapping account by Beuno and Colyvan (2011).

[5]See, for instance, Cassirer 1923, Duhem 1954, Goodman 1976 and 1978, Putnam 1987, and van Fraassen 2008.

[6]Alternatively, we could take the representation relation to be a primitive and leave it unanalyzed. There are important non-reductionist accounts of mathematical representation, for instance Suarez 2010, but a discussion of these cases would take us too far afield.

morphism is defined as a structure preserving map, or function, from one domain of mathematical structure to another, and van Fraassen correctly notes that "to define a function we need to have the domain and range identified first–and the question at issue [is] precisely how that can be done without presupposing that we already have a physical-mathematical relation on hand" (van Fraassen, 2008: 120).[7] If the definition of a morphism requires that a mathematical structure be defined on a physical relation, so that it can be representable in the mapping account, than we are faced with a dilemma: either the mapping account fails to account for applied mathematics, or the initial mathematization of the world is somehow already present. Berkovitz (2015) argues that the mapping account implicitly assumes that physical structure is mathematical, in a neo-Kantian or Pythagorean sense. If we ignore the problematic Pythagorean option,[8] we are once again led into a consideration of how mathematics is initially applied in the construction of our physical conception of the world.

If the scientific realist wants to base a theory of applied mathematics on the popular mapping account, then they need to clarify how mathematical concepts are brought to bear on the construction of our physical conception of the world.[9] As with any problem of conception, the realist needs to pay attention to where the points of convention lie. When we formulate certain scientific theories, especially in theoretical physics, mathematics plays an integral role in both the *definition* and *relation* of scientific concepts. The *definitional* role of mathematics delimits the domain of study by imposing a mathematical structure on the world. However, the *relational* role of mathematics provides the governing structure on this domain. The relationship between the definitional and relational roles of mathematics in complicated by the fact that mathematical concepts do not come free of charge. Implicit in

[7]The word "was" was substituted for "is" to reflect the tense of the discussion.

[8]If the Pythagorean view is associated with a naturalistic view of mathematics then it is subject to Brown's (2012) criticism and any rationalistic Pythagorean view seems to either collapse into the neo-Kantian view, or rely on an unaccounted for insight that borders on the mystical.

[9]This concern becomes more pressing when we consider whether or not a mapping-like account of applied mathematics is essential to any form of scientific realism. In the widely influential semantic account, scientific theories are thought to present structures or models that can be used to represent physical systems (Ladyman, 1998; 416). Any such account must clarify how mathematical models represent physical structures and it is difficult to see how the realist can account for the relation between a model and the world without either assuming that the representation relation is primitive, or presenting a mapping-like account of the representation.

their definition is a set of constraints that limit the types of physical structures to which they can be applied. These constraints are a direct result of certain assumptions that concern the underlying relations between the basic elements of a mathematical theory. Applying a particular mathematical concept to the physical world then entails that the basic assumptions in the underlying mathematical structure, such that the concepts may be well-defined, are satisfied by the world. The structural constraints implicit in the mathematical concepts dictate the type of physical phenomena that the theory can accommodate (Morrison, 2000; 109).[10]

If we are to untangle the web of issues related to mathematical representation, it is best to look to scientific practice and consider how a given mathematical theory comes to be applied. This paper will shed light on the essential conceptual pre-structuring of the world inherent in the application of mathematics by presenting an analysis of the use of the differential calculus in physical theory.[11] Specifically, this paper will treat the supposedly simple application of the differential calculus in the modern definition of Newton's second law. The application of the differential calculus requires that the world be pre-structured mathematically. This pre-structuring constrains the form of the world as understood within Newtonian theory. The constraints are a direct result of the formulation of the mathematical structure of the differential calculus. Our focus on mathematical constraint will highlight the dual role that mathematics plays in the definition and relation of physical concepts. The constraints imposed by the use of the differential calculus fall squarely within the purview of the definitional role of mathematics and, as such, delimit the applicability of the mapping account. This focus on the definitional and relational characteristics of applied mathematics will also showcase the role of convention and draw attention to the viability of any form of scientific realism that is based on the mapping account.

The body of this paper is comprised of three sections. The first section will develop the conceptual foundation of the differential calculus and identify the pre-structuring of the

[10] But here "type" should indicate form rather than kind.

[11] Note that the use of the term 'pre-structuring' should not be taken as a temporal relation but rather a necessary conceptual pre-structuring in the logical sense.

world inherent in its application. The second section will present a basic definition of Newton's second law and a discussion of the constraints that the differential calculus imposes on the structure of the world as conceived within Newtonian physics. This section will conclude with a discussion of what we take to be the limits of scientific realism, as conceived under the umbrella of the mapping account of applied mathematics. Finally, the third section will present a case study of a famous thought experiment by John Norton, simply called 'the dome'. The pre-structuring of the world required by the differential calculus offers a firm foundation for the mapping account of applied mathematics, but it also precludes certain physical structures from being understood within the confines of any theory based on the differential calculus. The modern formulation of Newton's second law is such a theory. The dome thought experiment provides a nice example of a hypothetical physical structure that fails to meet the necessary conditions for the differential calculus to be well-defined. Therefore, this structure is excluded by the pre-structuring of the world inherent in the application of the differential calculus. On the basis of this argument, we suggest that Norton incorrectly claims that the dome demonstrates the indeterministic nature of Newton's second law. Newton's second law actually cannot be applied in the thought experiment. This case study was chosen because it demonstrates the inherent danger in assuming that mathematics can be applied in a world of arbitrary structure.

The Conceptual Foundation of the Differential Calculus

The differential calculus plays an integral role in almost every theory of modern physics. In the modern formulation of Newtonian physics, it is constitutive of the very definition of motion. Formally, the differential calculus is applied to characterize the behaviour of a function in the infinitesimal neighbourhood of a point by providing a linear approximation to a function in that neighbourhood. But the differential calculus poses an interesting problem for any form of scientific realism based on mapping account of applied mathematics. The calculus

cannot be applied to an arbitrary function, but only to functions of a specific form. Therefore, the calculus can only be applied within a physical conception of the world in which the world is structured in a particular way. This pre-structuring of the world is not accounted for in the mapping account of applied mathematics and is a clear example of the mathematical construction implicit the in application of mathematics in the physical sciences. Our treatment of the differential calculus will begin with the definition of the concept of a function, and trace its development through the concepts of approximation, continuity, and the infinitesimal, culminating in a discussion of the differential and its role in the differential calculus.[12]

The conceptual foundation of the differential calculus begins with the notion of a function. A function is a relation, or map, from one domain of mathematical elements or structure to another. Functions are applied in the physical sciences to represent, among other things, entities (e.g. electrons and planets), constraint surfaces (e.g. the top of a table or a spacetime), and dynamical variables (e.g. force, position, and velocity) in the physical world. In each of these cases, a function serves to define the quantitative structure of the world by providing a map from some physical property, or structure, to a element, or structure, defined in $\mathbb{R}^n$, the n-dimensional space of real numbers. But how is a function is applicable to the world? Is this not the same question that lies at the heart of our discussion of the mapping account?

Within the conceptual system of a physical theory that is based on the differential calculus, the application of the concept of a function serves to define the initial mathematization of the world. This application of mathematics is itself a form of representation, but it is a representation akin to Goodman's characterization, in which we apply a representation to construe, classify, and interpret the world. In this sense, the application of the concept of a function serves to delimit the domain of study. But it is important to note that this initial

[12]In this section, We follow the development of the differential calculus provided by Loomis and Sternberg (1980). If the reader is familiar with the detailed formal development of the differential calculus, they may want to pass quickly through the mathematical parts of this section.

mathematization is both selective and productive. It is selective in that only those aspects of the world that are amenable to functional representation will enter into our physical conception of the world, e.g. extension and spatial-temporal location. When we conceive of the world as representable by functions, we limit our conception to only those aspects of the world that consistently allow such an interpretation. It is productive in that we fit the physical world for a "garb of ideas" to obtain an objective mathematical science (Husserl 1970; 54). The world as conceived through functions, is a quantitative mathematical world.[13]

We have barely gotten our feet wet, but the scientific realist might already feel slightly uneasy. If this initial mathematization of the world is a representation, in the sense of an interpretation, then there need not be any physical correlate to the mathematical structure. Rather, the mathematics is playing a definitional role that is constitutive of our physical conception. This issue is complicated by the fact that the definitional role of mathematics does not conclude with the application of the concept of a function, but rather only begins.

When we discuss functions in the differential calculus, we are usually interested in the behaviour of a function in the neighbourhood of a given point, but as we have already noted the differential calculus can only be applied to certain types of functions. In order to begin a discussion of the differential calculus, the functions we consider must satisfy four conditions:

> ***Condition 1:*** The function must be defined on at least one open neighbourhood of the point under consideration; except, maybe, the point itself.[14]
>
> ***Condition 2:*** The space of the function and the space of its domain must possess a norm (a definition of distance).[15]

[13]The scientific realist might protest that what is needed is not an exact quantitative world but only an approximation of the worlds inherent structure, however, any attempt to make the notion of approximation precise will have to provide a quantitative measure for the relation and, as such, would require an account of how this quantitative structure is defined and applied. This issue will be addressed in the next section.

[14]We allow for the possible exclusion of the point itself because, looking forward, the difference ratio of the calculus is not defined at the point under consideration.

[15]The concept of a norm allows us to provide a rigorous definition of distance and this provides us with a means to characterize an approximation and a coordinate system. In one dimension, it is customary to employ the absolute value of the difference between the elements, e.g. $|x - a|$, as the norm, but in multiple dimensions there are a few norms that work equally well.

Condition 3: The function must possess a limit in the neighbourhood of the point under consideration.[16]

Condition 4: The functions must be continuous.[17]

These conditions determine the form of the allowable physical structures that a theory based on the differential calculus can accommodate. These constraints are necessary for the concept of the infinitesimal to be well-defined and consistently applied. They represent the bare minimum that must be in place for our discussion of the differential calculus to begin.[18]

In the modern reformulation of the infinitesimal calculus, based on a rigorous foundation, infinitesimals are defined as functions that not only satisfy the previous four conditions, but also tend to zero as the element of their domain tends to zero, e.g. $\phi(t) \to 0$ as $t \to 0$. The difference ratio of the derivative is defined in terms of infinitesimals, $f'(x)$ is defined as $(f(x+h) - f(x))/t$ and this is simply the ratio of two infinitesimals (Loomis and Sternberg, 136). We usually say that the derivative $f'(x)$ exists and has a value a if $(f(x+h)-f(x))/t-a$ approaches 0 as $t \to 0$, or equivalently if $((f(x+h)-f(x))-at)/t$ approaches 0 as $t \to 0$ (Loomis and Sternberg, 136). In this case, $\phi(t) = (f(x+h)-f(x))-at$ "is an infinitesimal that approaches 0 *faster than* t (i.e., $\phi(t)/t \to 0$ as $t \to 0$)" (Loomis and Sternberg, 136). The fact that "ϕt converges to 0 faster than t as $t \to 0$ is exactly equivalent to the fact that the difference quotient of f converges to a" (Loomis and Sternberg, 137).[19] Therefore, the study of the derivative is equivalent to the study of the behaviour of

[16]In the ϵ,δ-definition of a limit, we say that a function $f(x)$ tends to a limit l as the element x approaches a if for every positive ϵ there exists a positive δ such that $0 < |x-a| < \delta \to |f(x) - l| < \epsilon$. It is important to note here that only functions possess a limit. Later on, when we discuss the differential calculus, keep in mind that the relation $(f(x+h) - f(x))/t$ expresses the ratio of two functions; we consider t to be a function not a variable.

[17]A function is 'continuous at a given point' if the limit, as defined above, exists at that point and the limiting value of the function, taken from the left and the right, is the same as the value of the function at that point. A function is 'continuous' if it is continuous at all points of its domain. more intuitive way to talk about continuity is through Hausdorff continuity. We say a set of elements is Hausdorff continuous if every pair of elements can be separated by an open neighbourhood.

[18]One could easily reformulate the following discussion in terms of continuity conditions, but we will base our treatment of the calculus on a discussion of infinitesimals, due to their intuitive appeal.

[19]Note: two commas were removed from the quote to fit the quote into the sentence structure.

infinitesimals.

Following Loomis and Sternberg, we may identify two special classes of infinitesimals: "big oh", O, and "little oh", o (Loomis and Sternberg, 136). A function falls under the class of "big oh", $f \in O$, if f is Lipschitz continuous at 0.[20] A function falls under the class "little oh", $f \in o$, if $f(x)/x \to 0$ as $x \to 0$.[21] Clearly, the numerator of the difference ratio of the derivative, written in the form $\phi(t) = (f(x+h) - f(x)) - at$, must be an infinitesimal of class "little oh". If this condition does not hold, the derivative cannot be well-defined.

The notion of an infinitesimal function defines an additional structure within our physical conception. We require that the functions we define on the world have a certain behaviour "in the small". But what type of constraint does this condition impose on the form of the functions we consider in the differential calculus? To answer this question, we will have to introduce the mathematical concept of a differential.

In our formal development of the concept of the differential, we will continue to follow Loomis and Sternberg (1980), and formally base the notion of a differential in terms of a general coordinate translation.[22] The coordinates of a function and its element are usually represented by an ordered pair containing the point under consideration and the value of the function at that point, $(a, f(a))$. We can always move the pair of elements to the origin by a coordinate translation of the form $s = f(x) - f(a)$ and $t = x - a$. In what follows, we will consider an ordered pair $(a, f(a))$ located near a point at which we would like to study the behaviour of a function. We can represent a general coordinate translation by the following diagram.

[20] a function is Lipschitz continuous if for all x sufficiently close to a, $|f(x) - l| \leq c|x - a|$, where l is the limit of the function and c is a constant.

[21] From this definition we can show that "big oh" is a subset of "little oh", $o \subset O$.

[22] A coordinate translation can represent a passive shift in the coordinate system, or an active translation that describes the motion of an object.

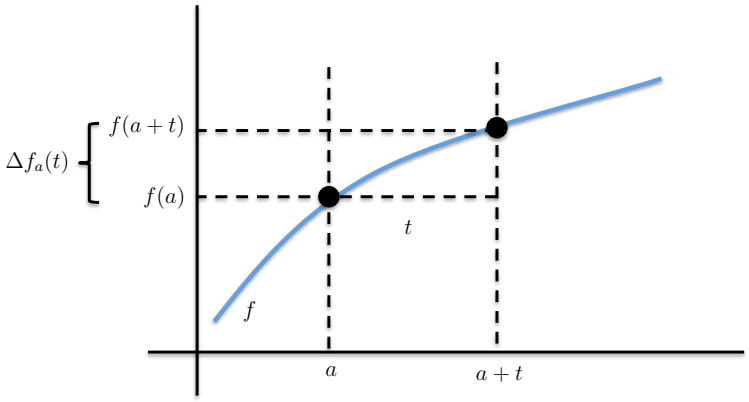


Figure 1: Diagram of a Coordinate Translation (Loomis and Sternberg, 141)

In the diagram it is clear that the image of f under the translation is given by the relation $\Delta f_a(t) = f(a+t) - f(a)$. $\Delta f_a(t)$ is simply the change in f brought about by the coordinate translation. The original curve, in the new coordinates, is the graph of $\Delta f_a(t)$.

We can now define the differential. In the new coordinate system, the equation for the tangent is given by the functional map $l(t) : t \to f'(a)t$; where $l(t)$ is the map from t onto the tangent (Loomis and Sternberg, 141). From this definition of $l(t)$, it is clear that the existence of the derivative $f'(a) = \Delta f_a(t)/t$ as $t \to 0$ is exactly equivalent to saying that $\Delta f_a(t) - l(t)/t \to 0$ as $t \to 0$ (Loomis and Sternberg, 141). Therefore, for the derivative to be well-defined in the infinitesimal neighbourhood of the point under consideration, the difference between the map $\Delta f_a(t)$ and the tangent $l(t)$ in that neighbourhood, given by $\Delta f_a(t) - l(t)$, must be an infinitesimal of the class "little oh", $\Delta f_a(t) - l(t) = o$ (Loomis and Sternberg, 141). To put the same point another way, we can say that the difference between $\Delta f_a(t)$ and $l(t)$ must tend to zero faster that t. It can also be shown that the expression $\Delta f_a(t) - l(t) = o$ is unique (Loomis and Sternberg, 141). The differential is defined as the "unique linear approximation $l(t)$ ... of f at a and is designated df_a" (Loomis and Sternberg, 141). From this definition, it is clear that without a well-defined differential, a derivative cannot be defined in the infinitesimal neighbourhood of the point under consideration.

The concept of a differential provides a valuable tool for analyzing the behaviour of a

function near a given point. In the limiting neighbourhood of the origin in the new coordinate system, the difference between an infinitesimal change in the function, $\Delta f_a(t)$, and the differential df_a is an infinitesimal of order o. The existence of a derivative at a given point requires that the infinitesimal behaviour of the function, $\Delta f_a(t)$, can be uniquely approximated by a differential map, up to an infinitesimal of order o. This means that the existence of the differential in the neighbourhood of the origin entails that the behaviour of the function in the neighbourhood can be approximated by a unique tangent. This can be seen clearly in the following diagram:

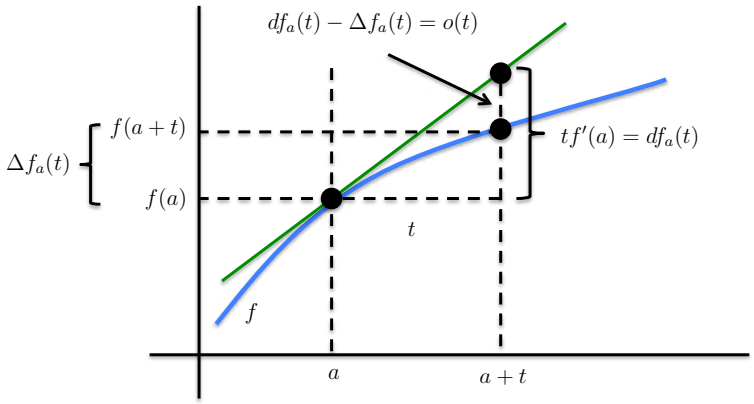


Figure 2: Diagram of a Coordinate Translation and the Differential (Loomis and Sternberg, 141)

The converse is also true. If the derivative does not exist, then the differential does not provide a unique approximation (up to class "little oh") of the function in the limiting neighbourhood of the point (Loomis and Sternberg, 146 -147). Therefore, the existence of a derivative entails the existence of a unique tangent that approximates the behaviour of the function up to class "little oh". Without this internal structure, the derivative cannot be defined. If we want to apply the differential calculus, then the functions we consider must possess this internal structure. This imposes a constraint on the form of any physical structure on which the differential calculus is applied.[23]

[23] But, here 'applied' should be read in the sense of a scientific realist's application of mathematics. Of course, one could apply the differential calculus to discrete systems by smoothing out the discontinuity through

We are now in a position to characterize the interpretation of the world that must be in place in order for the differential calculus to be applied. We can see that the functions that we apply within our scientific conception must be defined on at least one open neighbourhood of the point under consideration, possess a norm, possess a limit in the neighbourhood of the point under consideration, and be continuous, they must also possess a specific form such that the concepts and infinitesimal and differential can be well-defined and consistently applied. The scientific conception of the world, within any theory that applies the differential calculus, is defined and interpreted to possess this structure. When we interpret the world to possess a certain mathematical structure we at the same time construe and classify it. When we apply mathematical concepts to define the structure of the world, we project a mathematical structure onto the world in order to make it representable within the mapping account. And the choice among projectable mathematical concepts imposes a classification, which is simply a result of the governing mathematical conception that prevails within the larger theoretical structure.

The selective and productive interpretation of the world outlined in this section is based on the representation of the world given by the differential calculus, and is not based on any independent physical consideration. This presents a serious problem for the scientific realist, as the definitional role of mathematics does not necessarily possess a physical correlate. Rather, this mathematical pre-structuring of the world is a result of our intended representation the world given by the differential calculus. But this definitional role of mathematics is only half the story, and we now need to address the interrelation of mathematical concepts that takes place within a given physical theory.

idealization. But in this case the realist could no longer suggest that the mathematical structure of the differential calculus in any sense represents the structure of the world.

Newtonian Physics and the Differential Calculus

Newton's second law is simple enough to be familiar to almost every high school student and has remained a common discussion point in the philosophy of science. It expresses a relation between an impressed force on an object and the resulting change in the objects momentum. The modern definition of the law asserts that the force on an object is equal to the rate of change of the objects momentum, expressed as derivative of the momentum with respect to time. We write this symbolically as $\boldsymbol{F}(t) = d\boldsymbol{p}(t)/dt$. Within the mapping account, the scientific realist would want to claim that the differential relation maps a physical relation that holds between the physical force and the physical momentum of an object, which are represented by two time dependent vector functions $\boldsymbol{F}(t)$ and $\boldsymbol{p}(t)$, into $\mathbb{R}^3$, that is, if Newtonian theory were still accepted as a valid representation of the world.

But this simple narrative is untenable. We have noted that the application of the concept of a function is itself a representation, but one that serves to define the initial mathematization of the world. This mathematical construal, classification, and interpretation provides a quantitative structure to the world in order to provide a foundation for objective mathematical science. In this sense the interpretation of the world, as representable in $\mathbb{R}^3$, is both a selective and productive interpretation of the world that constitutes the basis of our physical conception. The use of functions delimits the conceptual system to only those aspects of the world that are amenable to functional representation. But what is more important in this case is the productive aspect of representation that fits the physical world for a "garb of ideas" to obtain an objective mathematical science (Husserl 1970; 54). For, the application of the concept of a function serves not only to define a quantitative structure on the world, but also to define which aspects of the world are to be represented as fundamental variables. Newton's second law produces an objective physical conception by setting a definition of inertial motion. Physical objects are thought to possess momentum, which remains constant unless a force acts on the object. Forces are construed to be a non-local relation that all objects enter

into as a result of their possession of certain properties, e.g. mass.

When we apply these two functions, $\boldsymbol{F}(t)$ and $\boldsymbol{p}(t)$, we do not represent force or momentum in the sense of a copy theory of representation, but produce a specific mathematical/physical conception of the world.[24] Newton's second law expresses a relation within this conception of the world, and if it meaningful at all, then it is a "law" of the world as representable within $\mathbb{R}^3$. But this law cannot be applied to arbitrary momentum and force functions, due to the constraints implicit in the definition of the differential calculus. Rather, we need to pre-structure our conception of the world such that the concepts of the differential calculus can be consistently applied and well-defined.

In order to apply the differential calculus, the function that represents the objects momentum must satisfy four conditions: namely, it must be defined on at least one open neighbourhood of the point under consideration, possess a norm, possess a limit in the neighbourhood of the point under consideration, and be continuous. The first condition is satisfied by stipulating that momentum, as construed within the Newtonian conception, is specified by a function that is defined on the neighbourhood of any point on its trajectory. The second condition is satisfied by imposing a Euclidean metric on $\mathbb{R}^3$.[25] The third and fourth conditions require that we represent the world in such a way that only continuous functions define the momentum of any object. Motion, as construed within the Newtonian conception of the world, is continuous, that is if we wish to apply Newton's second law.

The application of the differential calculus also requires that the function that represents the momentum of an object possess a certain internal structure. This structure is necessary so that the concept of an infinitesimal and differential can be consistently applied and well-defined. Specifically, what we require is that the functions possess a certain structure "in the small". The concept of a differential provides a valuable tool for characterizing the

[24] This claim is also supported by the existence of equivalent energy-based formalizations of classical mechanics.

[25] The Euclidean metric is defined as: $\|x\| = (\sum_{i=1}^{3} x_i^2)^{\frac{1}{2}}$. The absolute time of Newtonian physics is a one-dimensional space, and the absolute value function, $|x|$, provides a sufficient definition of distance in that space.

behaviour of a function near a given point. The existence of a derivative requires that the infinitesimal behaviour of the function can be uniquely approximated by a unique tangent. This in turn requires that the function can be uniquely approximated by a differential map in the infinitesimal neighbourhood of a given point. Motion, as construed within any conception of the world based on the differential calculus, is defined to have this internal structure "in the small".

However, the pre-structuring does not end here. In the modern formulation of Newtonian theory, the momentum function, $\boldsymbol{p}(t)$, is defined in terms of two other functions; one mass function, $m(t)$, and one velocity function, $\boldsymbol{v}(t)$. Formally, the momentum is defined as the mass times the velocity, $\boldsymbol{p}(t) = m(t)\boldsymbol{v}(t)$.[26] In the case, the differentiability of $\boldsymbol{p}(t)$ requires that both $m(t)$ and $\boldsymbol{v}(t)$ be differentiable. Therefore, the functions $m(t)$ and $\boldsymbol{v}(t)$ must also possess the necessary internal structure "in the small". And finally, the velocity function is defined as the rate of change of position in time, expressed as derivative of the position, $\boldsymbol{x}(t)$, with respect to time, $\boldsymbol{v}(t) = d\boldsymbol{x}(t)/dt$, and the position functions as well must possess the necessary internal structure "in the small" such that the concepts of the differential calculus can be consistently applied and well-defined. All of this pre-structuring must be in place in order to form the Newtonian conception of the world.

So where does this leave the scientific realist? On the one hand, we have a theory that is supposed to represent a physical relation that holds in the world. On the other, we have a set of mathematical definitions that construe, classify, and interpret the world in order to apply the theory. This initial representation of the world imparts it with a mathematical structure, and this pre-structuring undermines any form of scientific realism based on a copy theory of representation, such as the mapping account.

The fact is that any mathematical scientific theory that is taken to represent certain physical features of the world must address the implicit mathematization of the world. Husserl is right to note that "[m]athematics and mathematical science, as a garb of ideas, or garb

[26] Against usual convention, the mass functions is defined to be time dependent in order to highlight the fact that the continuity conditions apply equally to the mass and velocity functions.

of symbols of the systematic mathematical theories, encompassing everything which, for scientists and the educated generally, represent the life world, dresses it up as "objectively actual and true" nature" (Husserl, 1970; 54). Newton's second law only expresses a relation in the objective pre-structured world represented in $\mathbb{R}^3$. This mathematical law cannot copy, or map, a physical relation because there is no conception free physical relation that it can represent, as understood within a copy theory of representation.

Within the mapping account, It appears as though only a contingent form of scientific realism can be supported. Given a certain mathematical conception of the world, certain law-like relations hold, but these relations cannot be said to represent any innate structure in the world. At this point one might wonder if there any viable alternative open to the scientific realist. The answer will depend on whether or not the scientific realist can make do without a copy theory of mathematical representation. There is no question of whether or not Newton's second law expresses a functional relation, or map, from one domain of mathematical elements, or structure, to another. The mapping account provides an accurate description of the structure of the law itself, but this is not really the issue. The real issue relates to how a given mathematical structure, as a whole, represents a supposedly physical structure.

The real problem is that any symbolism, mathematical or not, harbours the curse of mediacy (Cassirer, 1946; 7). What is symbolized or represented is not a copy of what exists. The scientific realist might respond by abandoning the mapping account and noting that was is needed is not some exact copy of the world, which may indeed be impossible, but rather, a rough approximation to its structure. It may be the case that all this supposed pre-structuring is simply a form of abstraction or idealization that is typical of science in general, and in this case the real problem is that of abstraction and idealization, not of copying. The focus on approximation may change the nature of the question, but not its substance. The idea that mathematics might approximate, rather than copy, a physical structure still requires a clarification of how mathematics is brought to bear on the world. The concept

of approximation is equivocal, to say that a certain structure approximates another might indicate a closeness with respect to a given measure, an indication of similarity, the presence of common properties, or a number of other possible relations. The problem for the realist is to make the notion of approximation sufficiently precise without falling back onto the notion of approximately, or partially, copied structure. However, if the supposed 'closeness', 'similarity', or 'common property' is explicated in mathematical terms, then we end up right back where we started. The scientific realist needs to find a non-mathematical notion of approximation that is strong enough to support a viable realism but yet also weak enough to avoid the concerns associated with a copy theory of representation. Whether or not such an account can be found, the supposed marriage between the scientific realist and the mapping account is an unhappy one. Mathematics is not applied to map relations that hold within a physical system into an appropriate mathematical structure.

Although this paper has largely been a critical discussion of scientific realism and the mapping account of applied mathematics, it is not without practical importance. The pre-structuring inherent in the application of the differential calculus precludes certain structures from being well-defined within the confines of a Newtonian conception of the world. This pre-structuring limits the types of supposedly physical structures that the theory can accommodate. The form of the Newtonian conception of the world dictates the structure of the physical phenomena that the theory can accommodate, and in the final section, it is worthwhile to take a closer look at the constraints implicit in the application of Newton's second law.

The Newtonian Conception of Motion and a Case Study of Norton's Dome

What is the Newtonian conception of motion? The first thing we should note is that since Newton's second law can only be applied in the infinitesimal neighbourhood of a given

point, motion can only be defined up to an infinitesimal neighbourhood. If the realist wants to argue that Newton's second law is a true description of the phenomena, then it is a fuzzy description, and this inherent fuzziness is an unavoidable conclusion of the Newtonian conception of the world as governed by the differential calculus.

The description of motion, as defined by a Newtonian conception of the world, is complicated by the possibility of both constrained and unconstrained motion. In the case where the objects' motion is unconstrained, the objects' trajectory is defined solely with reference to the background space and time. If we consider a point-like object located at a particular point in space and time, we can discuss its motion with respect to a specified well-behaved force. At the initial time when the force is specified, the object is accelerated at a rate given by Newton's second law. The time evolution of the system can be specified by a unique trajectory in space and time. Since Newton's second law is an ordinary differential equation, and there are no additional constraints on the form of the position function, we can refer to the existence and uniqueness theorems of ordinary differential calculus in order to demonstrate that the trajectory of the object is indeed defined and unique (Kaplan,1973; 494-497). The structure of Newton's second law uniquely specifies the trajectory of an object with respect to the background space and time, and all of the conditions required for the differential calculus to be well-defined over every neighbourhood of each point along the trajectory are automatically satisfied.

The case of constrained motion is more complicated. A constraint imposes certain conditions on the form of an object's motion. For instance, we might consider the motion of an object constrained to the top of a billiards table. In this case, the motion of the object is constrained to the surface of the table and this imposes conditions on the form of the time evolution of the system given by Newton's second law. The problem that arises in the case of constrained motion is that the form of the constraint may impose undesirable conditions on the form of the function that defines the object's position. Since the differential calculus requires that this function possess a certain internal structure, only certain types of constraints

will allow for the differential calculus to be well-defined in the neighbourhood of every point. The good news is that many of the constraints that we consider within Newtonian theory are constructed to satisfy these conditions. The bad news is that there are many constraints that impose conditions on the form of a function such that it will possess neighbourhoods on which Newton's equation of motion simply cannot be applied. We will now discuss such a case.

Norton (2003) presents a now famous thought experiment simply called 'the dome', which attempts to demonstrate that Newton's second law allows for indeterministic solutions. Norton asks us to consider a point-like ball, of unit mass, located at the top of a frictionless, perfectly rigid, dome. The shape of the dome is given by $h = (2/3g)r^{\frac{3}{2}}$, where h is the height of the dome and r is the radial arc length measured along the surface of the dome (Norton, 2008; 787).[27] . The ball is subject only to the force of gravity.[28] At the start of this thought experiment, the ball is located at the apex of the dome.

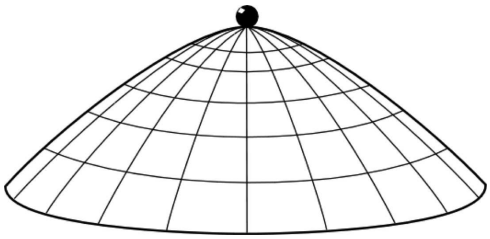


Figure 3: Norton's dome with static ball, taken from http://www.pitt.edu/ jdnorton/Goodies/Dome

Norton claims that the gravitational force, F, acting on the ball is given by: $F = (dh/dr) = r^{\frac{1}{2}}$ (Norton, 2008; 787). Since the ball has unit mass, Newton's second law states that the acceleration, $\frac{d^2r}{dt^2}$, is equal to this force, the result gives: $d^2r/dt^2 = r^{\frac{1}{2}}$. This is the equation of motion for a ball anywhere on the surface of the dome (Norton, 2008; 787).

We now come to the crux of Norton's argument. The equation of motion for a ball located at the apex is given by: $d^2r/dt^2 = 0$. One solution to this equation is: $r(t) = 0$

[27] From this point onwards the gravitational constant, g, will be set to 1

[28] Imagine that the dome is located within ahomogeneous gravitational field pointing downwards in the following diagram.

(Norton, 2008; 788). This is the solution that we naturally accept; that is, the radial coordinate of the ball remains constant. According to this criterion, the ball should not move. However, Norton claims that there is an alternative solution given by (Norton, 2008; 788):

$$r(t) = \begin{cases} \frac{1}{144}(t-T)^4 & \text{for t} \geq \text{T} \\ 0 & \text{for t} \leq \text{T}. \end{cases} \tag{1}$$

This solution states that the ball remains at the apex, for some arbitrary time, where it is subject to the equation of motion at the apex: $d^2r/dt^2 = 0$. However, spontaneously, the ball may begin to roll and is subsequently subject to the equation of motion for the surface of the dome: $d^2r/dt^2 = r^{\frac{1}{2}}$.

This solution states that the ball will remain at rest for some period of time, when $t \leq T$, and at $t = T$ the ball will spontaneously begin to roll down the dome. Notice the independence of these equations of motion on the radial direction the dome. If the ball is to move, there is no way of predicting the direction that it will go. Norton's conclusion is a result of the fact that the structure of the dome violates the Lipschitz condition and the associated existence and uniqueness theorem of ordinary differential calculus. It turns out that there is no way to predict at what time T the ball will begin to roll. If we add this fact to the independence of the equations of motion on the radial direction of descent, we observe a true indeterminacy in both the time and direction of descent. This is Norton's demonstration of indeterminacy at work in Newtonian physics. The whole situation is summed up nicely in the following diagram.

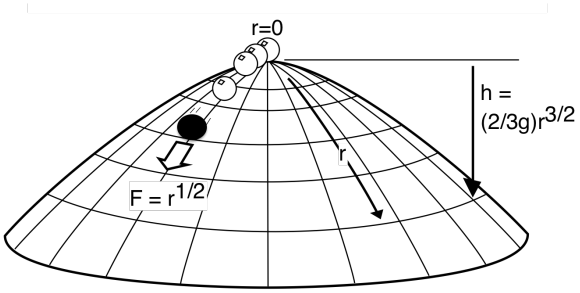


Figure 4: Norton's dome with falling ball, taken from http://www.pitt.edu/ jdnorton/Goodies/Dome

We intend to show that the problem with Norton's argument is that the differential calculus actually cannot be applied in the infinitesimal neighbourhood of the apex of the dome. This can be clearly seen if we consider the case of the ball rolling up the side of the dome towards the apex. We will show that as the ball approaches the infinitesimal neighbourhood of the apex, the differential structure breaks down. If the differential calculus cannot be defined in the infinitesimal neighbourhood surrounding the apex of the dome, then Norton cannot apply Newton's law in the infinitesimal neighbourhood of the apex, and the argument fails.

To begin, we can define a function, $\boldsymbol{x}(s)$ for the objects' position on the surface of the dome in terms of the arch length, s, measured from the apex. We can express the arc length, s, in terms of the time parameter, t, to define the objects position as a function of time, $\boldsymbol{x}(s(t))$. We will express the function $\boldsymbol{x}(s(t))$ in terms of a coordinate system fixed in the background Euclidean space. The origin of our coordinates will be centred on the apex of the dome with the apex itself occupying the point $(0, 1)$.

To simplify the problem, we can consider the case of a ball rolling up the right hand side of the dome.[29] In our coordinate system, the x coordinate of the dome is given by $x_1(s) = -\frac{2}{3}(1-s)^{\frac{3}{2}} + \frac{2}{3}$, and the y coordinate is given by $x_2(s) = 1 - \frac{2}{3}s^{\frac{3}{2}}$. The right half of the dome is then given by the equation $\boldsymbol{x}(s) = (x_1(s), x_2(s))$. The position of the ball along the dome as a function of time is then $\boldsymbol{x}(s(t)) = (x_1(s(t)), x_2(s(t)))$. Newton's second law states that

[29] We can define all of the other solutions from the radial symmetry.

$F = m\boldsymbol{x}''(s(t))$, where each prime indicates a derivative with respect to time, t. Expanding out the derivative by the chain rule, we find that $\boldsymbol{x}''(s(t)) = \ddot{\boldsymbol{x}}(s(t))s'(t) + \dot{\boldsymbol{x}}(s(t))s''(t)$, where each dot indicates a derivative with respect to arc length, s. We can immediately note that $\dot{\boldsymbol{x}}(s(t))$ is the tangent to the dome and $\ddot{\boldsymbol{x}}(s(t))$ is the normal to the dome. The behaviour of the derivative $\boldsymbol{x}''(s(t))$ in the infinitesimal neighbourhood of the apex is a function of the tangent and the normal to the dome. Therefore, we can get a good feel for how the ball will behave in the infinitesimal neighbourhood of the apex by studying the behaviour of the tangent and normal in that neighbourhood.

In our coordinate system, the tangent and the normal to Norton's dome are given by: $\dot{\boldsymbol{x}}(s(t)) = (\sqrt{(1-s)}, -\sqrt{s})$ and $\ddot{\boldsymbol{x}}(s(t)) = (-\frac{1}{2\sqrt{1-s}}, -\frac{1}{2\sqrt{s}})$, respectively. Immediately, we see that we are going to run into a problem. As the ball rolls towards the apex of the dome we see that the normal to the curve will blow up in the infinitesimal neighbourhood of the apex. This is simply a result of the fact that the curvature of the dome $\kappa(s) = \sqrt{(\ddot{\boldsymbol{x}}(s) \cdot \ddot{\boldsymbol{x}}(s))}$ blows up in the infinitesimal neighbourhood of the apex. Therefore, the derivative that represents the objects acceleration blows up as the object heads to the apex; Malament has come to the same conclusion (Malament, 2008). He suggests that the fact curvature blows up at the apex, shows that the apex of the dome has a zero fly-off speed, and might be considered to be a more of launching pad than a constraint surface (Malament, 2008; 13). Norton responded by noting that we could consider the ball, or in this case a bead, to be constrained to the surface of the dome by a perfectly rigid wire (Norton, 2008; 790). Norton claims that the wire would then provide the necessary constraint force to keep the ball on the surface of the dome, and Malament's concerns are easily alleviated. This apparent "solution" in no way alleviates Malament's concerns. The real problem is that the differential calculus simply cannot be applied in Norton's thought experiment.

Drawing from our discussion of the differential calculus, we can see what is going on. We know that the fact that the normal to the curve blows up in the infinitesimal neighbourhood of the apex indicates that the infinitesimal $\Delta\dot{\boldsymbol{x}}(s(t)) - d\dot{\boldsymbol{x}}(s(t)) = \dot{\boldsymbol{x}}(s(t+h)) -$

$\dot{\boldsymbol{x}}(s(t)) - l(s(t))$, where $l(s(t))$ is the tangent to the surface, is not of class "little oh". The issue is that $\Delta\dot{\boldsymbol{x}}(s(t))$ does not tend to its limit as fast as $h \to 0$. Therefore, when we take the derivative, we find that it blows up because the change in the function, $\Delta\dot{\boldsymbol{x}}(s(t))$, remains finite as $h \to 0$. If we cannot define an infinitesimal $\Delta\dot{\boldsymbol{x}}(s(t)) - d\dot{\boldsymbol{x}}(s(t))$ of class "little oh" then we cannot define a differential to the curve in the infinitesimal neighbourhood of the apex. If we cannot define a differential, then we cannot define a unique tangent, $l(s(t))$, to the curve, $\dot{\boldsymbol{x}}(s(t))$, that approximates the curve up to a class of "little oh". The problem is that we simply cannot determine the behaviour of the function $\dot{\boldsymbol{x}}(s(t))$ in the infinitesimal neighbourhood of the apex, because we cannot employ the concept of a differential to approximate the behaviour of the curve in that neighbourhood and if you cannot provide a unique linear approximation to the function $\dot{\boldsymbol{x}}(s(t))$ in the infinitesimal neighbourhood of the apex, then you cannot apply the differential calculus.

To get a feel for how pathological Norton's dome truly is, we can consider the normal force acting on the ball in the infinitesimal neighbourhood of the apex. The normal force on the ball over the surface of the dome is given by: $\boldsymbol{F}_{\perp}(s) = \sqrt{(1-s)}(-\sqrt{s}, -\sqrt{(1-s)})$; and its derivative is given by: $\dot{\boldsymbol{F}}_{\perp}(s) = (-\frac{1}{2}(1-2s)/(\sqrt{s}\sqrt{1-s}), 1)$. Right away, we see that the derivative blows up in the infinitesimal neighbourhood of the apex. Just as in our previous discussion, this indicates that we cannot define a differential to the force function that approximates the behaviour of the function in the infinitesimal neighbourhood of the apex. Therefore, we simply cannot define a well-behaved force acting on the ball in the infinitesimal neighbourhood of the apex. Geometrically, this is a result of the fact that the force swings though a finite angle in an infinitesimal neighbourhood.

The fundamental problem with Norton's thought experiment is that both of the functions employed in Newton's second law behave pathologically in the infinitesimal neighbourhood of the apex. All of this pathological behaviour is a simple result of applying Newtonian physics on a surface that is precluded by the pre-structuring of the world inherent in the Newtonian conception. Motion, as defined within the Newtonian conception of the world,

takes place within a mathematically pre-structured world that possesses a specific structure "in the small". This pre-structuring limits the form of the phenomena that the theory can describe.

Conclusion

Duhem was right to note that "[t]he role of the scientist is not limited to creating a clear and precise language in which to express concrete facts; rather, it is the case that the creation of this language presupposes the creation of a physical theory" (Duhem, 1954; 151). In the case of mathematics, the application of this language presupposes that our physical conception of the world has already been pre-structured mathematically. This initial mathematical pre-structuring of the world is a representation akin to Goodman's characterization, in which we apply a representation to construe, classify, and interpret the world. The choice among projectable mathematical concepts imposes a classification, which is simply a result of the governing mathematical conception that prevails within the larger theoretical structure. We saw that this initial mathematization is both selective and productive. It is selective in the sense that only those aspects of the world that are amenable to functional representation will enter into our physical conception of the world. And it is productive in the sense that we fit the physical world for a "garb of ideas" to obtain an objective mathematical science (Husserl 1970; 54). The world as conceived through mathematics, is a quantitative world. The real problem is that any symbolism, mathematical or not, harbours the curse of mediacy (Cassirer, 1946; 7). What is symbolized or represented is not a copy of what exists.

The mapping account of applied mathematics can only serve as a viable foundation for a contingent form of scientific realism. Given a certain mathematical conception of the world, certain law-like relations hold, but these relations cannot be said to represent any innate structure in the world. If a true scientific realism is to be grounded in a mathematical theory of the world, we must find an alternative to the mapping account of applied mathematics.

1 Bibliography

1) Batterman, Robert W. 2008. ”On the explanatory role of mathematics in empirical science.” The British Journal for the Philosophy of Science, Volume 61, No. 1. 1-25.

2) Berkovitz, Joseph. 2015. “The Propensity Interpretation of Probability: A Re-evaluation”. Erkenntnis, Volume 80, Issue 3. 629-711.

3) Brown, James Robert. 1999. *Philosophy of Mathematics.* London: Routledge.

4) Brown, James Robert. 2012. *Platonism, Naturalism, and Mathematical Knowledge*. New York: Routledge.

5) Cassirer, Ernst. 1923 *Substance and Function.* New York: Dover Publications.

6) Cassirer, Ernst. 1946. *Language and Myth.* New York: Dover Publications.

7) Goodman, Nelson. 1976. *Languages of Art.* Cambridge: Hackett Publishing Company.

8) Goodman, Nelson. 1978. *Ways of Worldmaking.* Cambridge: Hackett Publishing Company.

9) Husserl, Edmund. 1970. *The Crisis of European Sciences and Transcendental Phenomenology*. Evanston: Northwestern University Press.

10) Kaplan, Wilfred. 1984. *Advanced calculus, 3rd Edition.* Reading: Addison-Wesley.

11) Kitcher, Phillip. 1983. *The Nature of Mathematical Truth.* Oxford: Oxford University Press.

12) Ladyman, James. 1998. *What is Structural Realism?* Studies in the History and Philosophy of Science, Vol. 29, No. 3, pp. 409424.

13) Malament, David B. 2008. “Norton’s Slippery Slope.” Philosophy of Science, Volume 75 , Issue 5. 799 - 816.

14) Norton, John D . 2003. Causation as Folk Science, Philosophers Imprint 3 (4),

http://www.philosophersimprint.org/003004 Reprinted in H. Price and R. Corry, Causation and the Constitution of Reality. Oxford: Oxford University Press.

15) Norton, John D. 2008. "The Dome: An Unexpectedly Simple Failure of Determinism". Philosophy of Science, Volume 75 , Issue 5. 786 - 798.

16) Pincock, Christopher. 2007. "A Role for Mathematics in the Physical Sciences". Nous, Volume 41, Issue 2. 253275.

17)Pincock, Christopher. 2012. *Mathematics and Scientific Representation.* Oxford: Oxford University Press.

18) Putnam, Hilary. 1987. *The Many Faces of Realism.* LaSalle: Open Court.

19) Shapiro, Stewart. 1997. *Philosophy of Mathematics: Structure and Ontology.* Oxford: Oxford University Press.

20) Sternberg, Shlomo, and Lynn H. Loomis. 1980. *Advanced Calculus.* Reading: Addison-Wiley.

21) van Fraassen, Bas C. 2008 *Scientific Representation*. Oxford: Oxford University Press.

Against Selective Realism(s)

D. Tulodziecki[*†]

Draft/October 2016

Abstract: It has recently been suggested (for example, Lyons 2006) that realist responses to historical cases featured in pessimistic meta-inductions are not as successful as previously thought. In response, selective realists have updated the basic *divide et impera* strategy specifically to take such cases into account, and to argue, on this basis, that more modern realist accounts are immune to the historical challenge (cf. Vickers 2013). Using a case-study – that of the 19th century zymotic theory of disease – I argue that these updated proposals fail, and that even the most sophisticated recent realist accounts are just as vulnerable as their predecessors.

1 Introduction

The pessimistic meta-induction (PMI) targets the realist's claim that a theory's (approximate) truth is the best explanation for its success. It attempts to do so by undercutting the alleged connection between truth and success by arguing that highly successful, yet wildly false, theories are typical of the history of science.[1] There have been a number of prominent realist responses to the PMI, most notably those of Worrall (1989), Kitcher (1993), and Psillos (1999). All of these responses try to rehabilitate the connection between a theory's (approximate) truth and its success by attempting to show that there is some kind of continuity between earlier and later theories. One of the most widely discussed proposals has been Psillos's *divide et impera* strategy.

[*]tulodziecki@purdue.edu
[†]Department of Philosophy, Purdue University, West Lafayette, IN
[1]For a recent and new take on the PMI, see Frost-Arnold (2014).

Psillos argues (1999, Chapter 5), first, that realists ought to make the notion of a theory's success more stringent so as to include use-novel predictions, and, second, that realists face trouble only if it can be shown that those elements of a theory that "really fuel" that theory's genuine success are rejected and turn out to be completely false. It has recently been suggested – for example, by Lyons (2006) –, however, that Psillos's strategy is not as successful as previously thought. Lyons tests Psillos's move against a number of historical cases, and concludes that this "form of realism remains threatened by the historical argument that prompted it" (537). In response to Lyons, recent realists such as Vickers (2013) have argued that, once the selective realist strategy is updated appropriately, more modern realist accounts can, in fact, meet the challenge that Lyons has set.

In this paper, I argue that even recent, sophisticated realist accounts such as that of Vickers fail to be immune to the historical challenge and are just as vulnerable as their predecessors. I make my point by providing an example of such a case – that of the 19th century zymotic theory of disease, predecessor to the germ theory – and by carefully showing that this theory was highly successful in the realist's sense of 'genuine success'. I explain in detail what elements of the theory were responsible for its successes, by providing derivations of its predictions and the theoretical posits involved in making these predictions, and then show that the elements responsible for its success and that "really fueled" the relevant derivations were discarded in later theories and turned out to be completely false.

I will proceed as follows: In Section 2, I provide an overview of the zymotic theory of disease; in Section 3, I discuss its successes. Section 4 deals with the updated realist challenge and Section 5 is concerned with a derivation of the zymotic predictions and those posits of the zymotic theory that brought them about. In Section 6, I show how the updated realist challenge can be met, before concluding, in Section 7, that even the most sophisticated recent realist accounts are in trouble just as much as their predecessors.

2 The Zymotic Theory

The zymotic theory was the most sophisticated version of the miasma theory and dominated British disease theory from the 1840s to the 1870s (although it is frequently referenced well into the early 1900s). It sought to explain diseases in terms of complex interactions between miasmas and so-called

zymotic materials. Miasma was the result of rotting organic matter produced by decomposition processes. It would be dispensed in the air which, in turn, would act, via zymotic principles, on individual constitutions, causing one of several diseases (cholera, yellow fever, typhus, etc.), depending on a number of more specific factors. Some of these were thought to be extraneous, such as weather, climate, and humidity, and would affect the nature of the miasmas themselves; others were related directly to the potential victims and thought to render them more or less susceptible to disease. Lastly, there were a variety of local conditions that could exacerbate the course and severity of the disease, such as overcrowding and bad ventilation.

The term 'zymotic' (from the Greek for fermentation) goes back to William Farr (1807–1883), Statistical Superintendent of the General Register Office from 1842 to 1879. Farr coined this term to indicate that disease processes "are of a chemical nature, and analogous to fermentation; by which they are moreover to a certain extent explained" (1842, 201), yet not identical to vinous fermentation. Since decomposition figured heavily in the various accounts of disease causation, disease theorists drew heavily on contemporary chemical theories, such as those of Liebig, who had both a comprehensive system for explaining the various morbid processes of decomposition, putrefaction, and fermentation, but also his own specific zymotic pathology. Chemical theories like those of Liebig were well suited to explaining diseases, because they explained the interaction between living and non-living things, such as human bodies and the environment, and they did this on a molecular basis. Moreover, Liebig's chemical theories were popular, highly respectable, and they had already had great successes, and so the zymotic theory may be seen as drawing on some of the most successful science at the time.

Liebig and Farr held a so-called contact theory of decomposition. According to this, diseases occur as a result of introducing into the body (through inhalation or direct contact) various zymotic principles. These were thought to be "organic matter in a state of pathological transformation" (Farr 1842, 202). This would be absorbed by the blood, and, through the transformation, zymotic diseases had "the property of communicating their action [i.e. decomposition], and effecting analogous transformations in other bodies" (ibid.). The zymotic principles were the 'exciters' of the various diseases and "in the blood corresponding bodies exist, which are destroyed, and by the transformation of which the exciters are generated or reproduced" (ibid., 199). In short, pre-existing stuff in a victim's blood catches the process of decomposition and communicates this state to other particles of blood (which,

in turn, would transmit it to various body parts), until it ran out of susceptible particles to contaminate. The underlying idea of the contact theory was that zymotic matter was like ferment, a volatile chemical substance that could transfer its volatility to other materials. So, just as ferment produced fermentation, zymotic material produced disease (zymosis).

Two things about the zymotic theory are worth stressing: first, zymotic material was not a specific substance; according to the zymotic theory, the disease was not the (presence of) zymotic materials, but the zymotic processes of transformation. Second, the zymotic account was purely chemical, and Liebig (and others) explicitly rejected the view that zymotic materials were living organisms.

3 Successes of the Zymotic Theory

The zymotic theory was highly successful with respect to a number of phenomena. Specifically, it was successful both in terms of explanation and prediction. Among its explanatory successes were explanations of well-known disease phenomena, such as the fact that diseases were known to be seasonal, and often tied to particular regions (such as marshy ones) or particular locations (barracks, prisons, etc.). Similarly, it was well known that sickness and mortality rates in poor, crowded urban centres were worse than in their less poor and crowded counterparts, and that, in turn, those parts were affected worse than rural areas. It could also explain a number of facts tied to epidemics, such as why epidemics began, took the course they did, and then subsided, yet often came back several years later. It could account for the fact that epidemic diseases often moved around when there was no known contact with previous victims, why quarantines were ineffective for some diseases (such as cholera), why only some, but not all people were affected by a given disease, and, lastly, why certain diseases were endemic.

The zymotic theory could explain all of these phenomena through its claims that decomposing material produced miasmas. Diseases peaked when conditions for putrefaction were particularly favourable: this was the reason why certain diseases were particularly bad during periods of high temperature and in certain geographical regions (for example, the many fevers in Africa), why urban centres were much more affected than rural areas, and why even specific locations in otherwise more or less healthy areas could be struck (sewage, refuse, and general 'filth' would sit around in badly ventilated areas). More-

over, since zymotic material interacted with the blood, the zymotic theory could provide an account for individual disease susceptibility, and explain why certain diseases were contracted only once (once the relevant material in a victim's blood had been 'converted', the person became immune).

While this degree of explanatory success is quite impressive, as we have seen, however, realists tend to think that, in addition, genuine success also requires a theory to make use-novel predictions, i.e. predictions that did not play a role in the theory's original formulation. Here, the zymotic theory also delivers. It predicted, for example,

(i) that the air in areas with higher disease incidence ought to be worse than the air in healthier areas; more specifically, that it should contain more decomposing organic material,

(ii) a number of different disease incidence patters, based on the prevalence of decomposing and putrefying materials (in conjunction with facts about ventilation), including

  (a) relationships between disease prevalence and season, temperature, rainfall, wind, and so on,

  (b) a relationship between disease incidence and population density,

  (c) a relationship between disease incidence and elevation,

(iii) facts about the course and duration of various epidemics,

(iv) facts about the relation between mortality rates and different occupations, and, lastly,

(v) facts about relationships between mortality from various diseases and age.

Thus, the zymotic theory had successes on both explanatory and predictive levels. Before showing (in Section 6) that these successes were, in fact, due to a number of essential working posits that did not get retained in successor theories and that turned out to be completely false, however, let us be clear about what exactly the new selective realist challenge amounts to.

4 The New Selective Realist Challenge

In response to the pessimistic meta-induction, besides making the notion of success more stringent, selective realists have suggested that we ought to focus only on those parts of past theories that were, in fact, responsible for their genuine success. Prominent approaches include Worrall's structural realism, Kitcher's distinction between working and presuppositional posits, and Psillos's *divide et impera* strategy. And, while this line of response has also been popular among more recent realists, such as Harker (2013), Peters (2014), Saatsi (2005), and Vickers (2013), they also acknowledge the shortcomings of the traditional responses. In this vein, Vickers, for example, argues that the basic *divide et impera* strategy needs to be updated, since (i) first, there now are cases of successful theories that did make novel predictions but that, nevertheless, turned out to be completely false (cf. Lyons 2006), and (ii) second, "the *divide et impera* position needs significant refinement, especially concerning the crucial concepts *scientific success* and *responsible for* on which so much weight is placed" (2013, 190).

Thus, Vickers agrees with his predecessors that a given historical case poses a problem only "if posits that 'really fuel the derivation' [of a novel prediction] turn out to be definitely not approximately true" (194), but he thinks more light needs to be shed on what it means for a posit to "really fuel a derivation". To this effect, Vickers proposes a distinction between derivation-external and derivation-internal posits. The former are those that "merely influenced the thinking of scientists" (198), but, since they only guide scientists and are not part of the relevant derivation, they are not eligible for realist commitment. The latter are those "posits [that] were actually involved in the derivation of that prediction" (198). However, Vickers argues, derivation-internal posits are not "necessarily the 'working posits' since any individual posit might 'contain within it' some other posit that is the real working part" (198). In other words, a posit might contain a 'weaker' posit that is sufficient for the prediction, such as the posit "the passengers are 50kg too heavy" containing within it the weaker posit "the passengers are too heavy" (198.; originally due to Saatsi 2005, 532).

Thus, contrary to previous realists – especially Psillos, according to whom scientists' judgements play an important role in determining what fuelled a derivation – Vickers thinks that realists ought to commit themselves only to posits that do logical work, and moreover, to their weakest possible versions (i.e. if one weakened the posits any more, one could no longer derive the

prediction). Thus, Vickers makes a distinction between actual, historical derivations, and (possibly much weaker) logical, epistemic derivations (cf. Peters 2014, 386). And, while Vickers believes there are now cases showing that there are theories that contain working posits that turned out to be false, he does not think that there are cases showing that there are essential parts of derivation-internal posits that turned out to be false. Coming up with such cases, then, is the new and updated selective realist challenge.

5 Zymotic Predictions

In order to show that the zymotic theory can rise to it, let's look at some of its predictions in more detail, starting with its predictions concerning air quality. Here, the zymotic theory predicts, (i) first, that air quality ought to be proportional to disease incidence, so that the 'right' locations should have good and bad air, respectively, and (ii) second that differences in air quality ought to be related to decomposition and ventilation. A number of mid-19th century chemists tried to test these predictions; however, for brevity's sake, I will restrict my focus to a small subset of the experiments of Robert Angus Smith (1817–1884), a contemporary of Farr's, and often cited by the latter.[2]

Smith began by collecting indoor condensation liquid from crowded rooms and compared it to fresh rainwater, finding that the indoor liquid, but not the rainwater, had a strong perspiration smell, and, "on standing it formed a glutinous mass in which the microscope revealed "Confervae", "greenish globules", "various species of Volvox" [a type of algae], and "monades many times smaller"" (ibid., 219–220; Smith, 1848, 18). While this was not a strong result by itself, Smith believed that it at least showed that the the indoor liquid contained organic material on which the Volvox and the monades were feeding (ibid., 220). Further, upon burning the liquid's residue, he obtained the smell of ammonia, which was significant, since ammonia was tied to the last stage in decomposition. At the larger scale of towns, these results were thought to be exacerbated, not just because of the various exhalations of living bodies, but, in addition, those of animal refuse and fuel combustion (Smith, 1848).

Smith also tried to measure air quality more directly. He washed air samples by changing the air in a bottle containing distilled water, shaking the

[2]Smith was not a particularly distinguished scientist (Eyler 1980, 216), but is, for this reason, quite representative of a number of people and their work.

bottle after each change of air (with up to 200 changes), and then performed a chemical analysis on the water, determining its amount of organic material. He found these results to confirm those previously obtained from the experiments above (Smith 1859, 218–225; cf. also Eyler, 220).

Most telling, however, were experiments in which he tried to show that there was organic material in vapours given off by putrefying meat, blood (due to the strong smell, he only performed a limited number of these), and cesspools. He proceeded to compare these samples to the air in a number of different locations – everything from different areas of Manchester, to "closely packed railway carriages[s]" (1859, 221), to the air in bedrooms before and after someone had slept in them, to the occasion on which a "strong smell of a sewer entered my laboratory" (221), the fronts and backs of various houses, alleys, and so on. Smith obtained rather a large variation in air quality, and, specifically, concluded that by inhaling putrid air from decomposing animal matter (such as decomposing mutton), "we might be inhaling 9000 times more of some organic substance or other than we should be doing by inhaling the purest air" (1859, 222). These figures include the vapours given off by the putrefying meat, but, without these, he concluded that the difference among different areas of town is about 22 times, and, within industrial areas, ranges from about 9 to 22 times, while over the Atlantic Ocean and the Highlands, the air was clean (223; cf. also Eyler 220–221). Most importantly, however, as Eyler points out, the ranges Smith found were in proportion "to the range of the crude death rates for the districts of Manchester" (221): in the districts with the highest death rates, the air had most organic material in it, and the ones with the lowest death rate had the cleanest air. Smith concludes that "[t]hese differences ... enable us to account for the number which represent the deaths of the various districts" (1859, 223).

He also performed an investigation for the mining commission, during which he observed people in air-tight lead chambers, recording their pulse, respiration, and so on. He systematically measured the carbonic acid concentration inside the chamber, and concluded that carbonic acid was quite harmful, that it "almost always comes in bad company" (Eyler, 222), and then proceeded to use carbonic acid concentrations to test how well or badly ventilated a given place was.

However, Angus Smith's experiments are not the only ones speaking in favour of the zymotic theory. As we have already seen, the zymotic theory also predicted a number of other relationships, such as those between disease incidence and population density and between disease incidence and distance

from sources of decomposition. Both of the latter were confirmed by Farr. First, Farr showed that "the mortality of town districts has a certain relationship to their density" (207). Based on his analyses of the data sets of a number of Annual Reports of the General Register Office, Farr came up with a law on whose basis he made a number of precise predictions. When he then proceeded to compare the calculated, expected mortality to that actually observed in the 30 London statistical districts, he found the results agreeing "very directly with the results of direct observation" (ibid.; due to space constraints, I won't discuss this result in any detail, but, for a quick flavour of the kind of law Farr put forth, see figure 1 (ibid.)).

Figure 1: Farr's Density Equation

Given the mortality of St. James's district = ·02145; the density of the population 145059 to a square mile; the density of St. George's, Hanover-square, 39018 to a square mile, what was the mortality of St. George's, Hanover-square?

Here $\sqrt[6]{\frac{d}{d}\,m'} = m$; and substituting the figures in the formula, the result agrees very exactly with the results of direct observation $\sqrt[6]{\frac{39018}{145059}} \times ·02145 = ·0172$, the mortality of St. George, Hanover-square, the mortality observed having been ·0171.

The mortality of the city of London deduced from the mortality of St. George, Hanover-square, and the densities of the two districts is

$$\sqrt[6]{\frac{d'}{d}\,m} = m'; \text{ or } \sqrt[6]{\frac{94488}{39018}} \times ·01707 = ·0198.$$

The density of a district is deducible from the same formula. It may be expressed, however, differently; namely, by the number of square yards to a person, and denoted by y' in the district where the number of square yards to a person is greatest: then

$\frac{m}{m'} = \frac{\sqrt[6]{y}}{\sqrt[6]{y'}}$; and $\left(\frac{m}{m'}\right)^6 = \frac{y}{y'}$; consequently, $y = \left(\frac{m}{m'}\right)^6 y' =$ the number of square yards to a person in the least dense district.

His crowning achievement, however, was his elevation law, relating cholera mortality to soil elevation. This relation relation followed straightforwardly from those parts of the zymotic theory having to do with the dilution of miasma in the atmosphere. Here, what Farr did was to capture the exact relation between the decline of cholera and increased soil elevation in the form of the following equation: $c = c' \times (e' + a)/(e + a)$($e$ is the elevation above the Thames high water mark, c the average cholera mortality rate at that elevation, and e' and c' the elevation and mortality at a higher elevation, a is a constant; 1852, lxiii). Farr then calculated the expected series according to the formula, compared it to the actual series recorded in London, and found remarkable agreement (1852, lxiii). Seeking further confirmation, he

then proceeded to "submit the principle to another test, by comparing the elevation and the mortality from cholera of *each sub-district*", and found that this "entirely confirms the announced law" (xv-xvi). Lastly, Farr's predictions were also confirmed by others in different regions. For example, "William Duncan, Medical Officer of Health for Liverpool, wrote that when he grouped the districts of his city by elevation as Farr had done, that cholera mortality in the last epidemic obeyed Farr's elevation law for Liverpool as well" (Eyler, 1979, 228).

6 Meeting the New Realist Challenge

Note that all of the above predictions were use-novel: they could not have played a role in the construction of the zymotic theory, since they were not even formulated by then, and, in Farr's case, the data on which his laws were based did not even exist. As such, the case of the zymotic theory meets the realists' criteria for 'genuine success'.

However, as we have seen, Vickers does not think that this is enough; in addition, he requires that it be shown that the derivation of these predictions involved ineliminable parts of derivation-internal posits that cannot be weakened. Here, I want to argue that the zymotic theory can meet this challenge, too.

The first question is what the zymotic theory's relevant posits are. Crucial to the derivations of the above predictions are the following three:

1. Organic matter given off by putrefying material is the exciting cause of diseases.

2. This organic matter is suspended in the air (i.e. there are miasmas).

3. This, in turn, is transmitted through the air.

Clearly, all of the above posits are derivation-internal, not derivation-external: they did not just guide practitioners' thinking, but they were all crucially involved in making the various predictions. They were essential in the air quality predictions confirmed by Angus Smith: if any of the above posits are taken away, the entire prediction disappears. While the second and third are directly about quality, the first is also necessary: without specific reference to (sources of) decomposition, we lose the geographical and

other location patterns, and, as a result, the entire prediction about air differences. Thus, all of these posits are clearly doing work in the production of the prediction. Similarly for Farr's predictions about density and elevation. Both putrefying material as a source of miasma and air as a medium of its transmission are crucial, since, taking either one away makes the prediction disappear. Thus, it is clear that the above posits are doing work.

However, as we have seen, according to Vickers, this is not enough – it also needs to be the case that they don't entail weaker versions of themselves, on the basis of which the prediction still goes through. Examining our three posits, we see that this, too, is the case. Ironically, the fact that the zymotists knew relatively little about the chemical make-up of the alleged miasmas is to their advantage here, since, absent any specific information about the make-up of the allegedly responsible organic matter, these posits are already as weak as they can get: they effectively state that *whatever* is given off during decomposition is transmitted through the air and involved in causing diseases. Weakening them any more would take away either the decomposing sources, or air as a medium, and, as we have already seen, both of these are necessary in order to make the predictions in the first place. Thus, since these posits cannot be weakened, they ought to count as essential parts of derivation-internal posits, according to Vickers. Moreover, they did not get retained in any way, shape, or form, in the germ theory: decomposition is not responsible for disease, neither is disease material dispensed in the air, neither is the air a medium for disease transmission. But, if that is right, according to Vickers, and by his own admission, we ought to have been realists about miasmas.[3]

So much for miasmas, but what about zymes? Further zymotic posits we might add to the miasmatic ones above are:

1. Diseases are a type of decomposition,
2. Disease processes are analogous to processes of fermentation (which is a type of decomposition),
3. Zymotic material acts like a ferment on pre-existing stuff in a victim's blood.

[3]Note that this poses a problem not just for Vickers, but for selective realists more generally.

Now, clearly, none of these posits are necessary for the predictions we have seen, since it was possible to provide derivations of these predictions without appealing to any of the zymotic posits. Thus, obviously, they are non-essential. However, I want to stress at this point that, even if the zymotic posits fail to be essential, note that the miasmatic posits clearly are, and that this is already enough of a problem for selective realists. After all, selective realists face trouble if essential parts of derivation-internal posits turn out to have been completely false, and the miasmatic posits fit that bill. But, of course, it is not a requirement that every (rejected) posit be essential (which is obviously not right). Thus, showing that the miasmatic posits are essential and were rejected suffices to get the realist into trouble.

More interesting, however, is the fact that even though the zymotic posits might not have been essential for making predictions, they play a different, and perhaps equally important role in the zymotic theory: that of providing unifying explanations.[4] More specifically, among other things, the zymotic posits explain why there is a link between decomposition and disease (they are essentially variations of each other), why diseases vary with season and temperature (it was well known that fermentation processes are temperature-sensitive), why certain diseases only occurred once (victims' blood would run out of material to ferment), why certain diseases were childhood diseases (children's blood was different from adults'), and why epidemics began (existence of sufficiently virulent zymotic material) and ended (lack of new victims with the appropriate blood). Lastly, it had been clear for some time that disease material needed to replicate itself somehow, and fermentation processes offered an explanation of how this was possible. The various zymotic posits above can make sense of all of these at once. And, while there are no concrete predictions based directly on these posits, it is clear that the above explanations disappear without zymosis. Thus, while zymotic material might not have been primarily responsible for the theory's predictive success, it was certainly crucially implicated in the theory's explanatory success. Without the fermentation aspect of zymosis, the entire disease mechanism would have disappeared, and with it disease pathology. Moreover, without it, the link between diseases and decomposition would have been lost, and, as we have seen, without decomposition, the predictions of the zymotic theory disap-

[4]Peters (2014) stresses the importance of this in accounts of essentialness. For this reason, the zymotic theory might also turn out to be problematic for Peters' proposal. However, due to space constraints, I will not pursue this further here.

pear, too.[5]

7 Conclusion

What, then, is the upshot of this? What is clear is that the zymotic theory made use-novel predictions, such as those about air quality, and Farr's predictions about elevation and density. Miasmatic posits were essential derivation-internal working posits that, further, could not be weakened, and, so, according to Vickers, they deserved realist commitment. Yet, they were rejected, and no miasmatic posits, as we have seen, were carried over to the germ theory. Zymotic posits were not involved in the predictions in these crucial ways, but were essential to the theory's explanatory power, since, taking away zymosis left a disease theory without a disease mechanism, without a pathology, and without any explanation of many of the well-known disease phenomena that needed accounting for, including the all-important link to decomposition. Thus, even modern, sophisticated selective realist accounts, such as that of Vickers, cannot rise to the historical challenge that underlies the pessimistic meta-induction. Realism remains in as much trouble as ever, at least on this front.

8 References

Eyler, John M. (1979). *Victorian social medicine: the ideas and methods of William Farr*. Johns Hopkins University Press.

Eyler, John M. (1980). "The Conversion of Angus Smith". *Bulletin of the History of Medicine*, 54(2), 216.

Farr, William (1842). "Letter", *4th Annual Report to the Registrar General, Appendix*, London: W. Clowes

Farr, William. (1852). *Report on the Mortality of Cholera in England, 1848-49*, London: W. Clowes.

[5]For brevity, I cannot discuss in detail why zymes cannot be regarded as proto-germs. I want to stress, however, that even if such an argument could be made (although I am confident that it couldn't), this does not help the realist, since the miasmatic posits were the ones responsible for the zymotic theory's novel predictions, and miasma certainly did not get retained in the germ theory.

Frost-Arnold, Greg (2014). "Can the Pessimistic Induction be Saved from Semantic Anti-Realism about Scientific Theory?" *British Journal for the Philosophy of Science* 65(3): 521-548.

Harker, David (2013). "How to Split a Theory: Defending Selective Realism and Convergence without Proximity". *British Journal for the Philosophy of Science*, 64(1): 79-106.

Kitcher, Philip (1993). *The Advancement of Science*. Oxford University Press

Lyons, Timothy D. (2006). "Scientific realism and the stratagema de divide et impera". *British Journal for the Philosophy of Science* 57(3):537-560.

Peters, Dean (2014). "What Elements of Successful Scientific Theories Are the Correct Targets for "Selective" Scientific Realism?". *Philosophy of Science* 81(3): 377-397.

Psillos, Stathis (1999). *Scientific Realism: How Science Tracks Truth*. Routledge

Saatsi, Juha (2005). "Reconsidering the Fresnel-Maxwell Case Study". *Studies in History and Philosophy of Science*, 36: 50–38.

Smith, Robert A. (1848). "On the air and water of towns", *Report B.A.A.S.* 18: 16–31.

Smith, Robert A. (1859). "On the air of towns". *Q. J. Chem. Soc. Lond.*. 11: 196–235.

Vickers, Peter (2013). "A Confrontation of Convergent Realism". *Philosophy of Science* 80(2): 189–211.

Worrall, John (1989). "Structural realism: The best of both worlds?" *Dialectica*, 43(1-2): 99–124.