Philosophers' Imprint

volume 19, no. 13
april 2019

Method Pluralism, Method Mismatch, & Method Bias

Adrian Currie & Shahar Avin

Centre for the Study of Existential Risk, University of Cambridge

Acknowledgements

This paper was discussed by the social epistemology reading group in Cambridge, who provided extremely helpful feedback, as did Sabina Leonelli, Johannes Himmelreich, and two anonymous reviewers. The ideas were presented both at the University of Sheffield and at the European Philosophy of Science Association meeting in Exeter; we're grateful for the help both audiences provided. This publication was made possible through the support of a grant from Templeton World Charity Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation.

1. Introduction

Philosophers, historians, and sociologists of science have long agreed that there is no single scientific method. How to best generate knowledge depends crucially on the kind of knowledge in question, as well as technological, social, and theoretical context, and the kind of system the investigation targets: what we'll call the "epistemic situation".[1] Accepting that good knowledge-production admits of a plurality of methods, and that these are more or less appropriate in different epistemic situations, leads us to the possibility of method *mismatch*. The adopted method might be inappropriate for the epistemic task. Method mismatch likely occurs due to method *bias*: tacit or explicit ideas about knowledge-production might influence scientific practice. Our target, then, is how opinions about good method in a scientific community — as reflected in publishing practices, for instance — might influence the nature and productivity of that community.

In this paper, we articulate a form of method plurality, allowing us to explore both mismatch and bias *vis-à-vis* method. We'll understand method plurality by distinguishing between two properties of evidence. First, what we'll call *sharpness*: how incisive evidence is regarding a hypothesis. Second, what we'll call *independence*: the amount of overlap between the background theory which underwrites different

[1] For fuller discussion of this notion, see Leonelli 2016, Currie 2018.

evidence. In light of this, we'll characterize two scientific strategies, one targeting sharpness — methodological "obligates" — and the other targeting independence: methodological "omnivores". We'll then consider under what conditions method mismatch could occur, exploring two aspects of epistemic situations: first, the evidential context investigators face; second, community-level preferences for different evidence types.

We'll not claim that method bias is necessarily problematic: indeed, as we'll show, sometimes introducing it can ensure a diversity of strategies are employed. Moreover, given that method bias is, in effect, a community's preference for types of evidence, it is often unavoidable. The take-home message is that the make-up of a scientific community depends both upon evidential context, and upon what is recognised as "good evidence" or practice in that community. Publishing standards, funding decisions, the opinions of referees, and so forth affect what kind of work gets done. Recognising that, on the one hand, different investigative strategies are more appropriate in different contexts, while on the other hand, opinions about what good science looks like might either follow or work against those contexts, allows us to see the kinds of effects method bias might have. Moreover, this itself underwrites an argument that understanding scientific evidence without taking the relevant community's views about good evidence, and the epistemic situation, into account, is wrong-headed. Indeed: we'll argue that consideration of such community properties is required for understanding evidence.

In addition to these points about the social epistemology of science, we also take ourselves to be making a contribution to discussion of how we should understand and make optimal epistemic progress, under constraint, in a Bayesian framework. This more technical discussion largely plays out in the footnotes, specifically footnotes 8, 12, 18, 20, and 21.[2]

We'll analyse and explore method plurality, mismatch, and bias, via agent-based modelling, which we introduce and utilize in sections 3 through 5. We construct an epistemic landscape which represents the two properties of evidence which interest us. We then add agents which adopt the strategies we've mentioned, and publishing standards which reflect the community's method bias. We'll establish a link between properties of the landscape and the success of different strategies. Some landscapes favour obligates, others omnivores. We'll then explore the effects of method bias by considering to what extent changing publishing standards can influence the kind of evidence which is generated in the population, and the make-up of the population in terms of strategies.

According to our model, method bias can maintain a diversity of strategies and evidence in a community. This can occur even when the landscape heavily favours one strategy over another. Moreover, making it easier for the disadvantaged strategy is more effective than making it harder for the favoured strategy. Finally, the trade-off between method bias and a community's productivity is not simple: often gains in one arena are not equally matched by losses in another. This suggests that, under some conditions, maintaining a diversity of scientific strategies can actually increase the productivity of a community overall. Before describing our simulation work, in section 2 we'll analyse sharpness and independence.

Our primary goal in this paper, then, is to identify and analyse an otherwise unnoticed kind of bias whose recognition is made possible through a commitment to method pluralism and a recognition that some methods are more useful in some contexts than others. In doing so, we also make original use of "epistemic landscape" models, both in the dynamics they instantiate and in how we interpret them. We'll finish in section 6 by discussing circumstances in which method bias might be particularly egregious, and by arguing that philosophical analysis which approaches evidence abstracted from aspects of epistemic situations — community standards in particular — are inadequate for many instances of knowledge generation in science.

[2] We thank an anonymous referee for pointing out this angle, and for helpful suggestions on how to make it explicit.

2. Sharpness & Independence

In this section we provide conceptual underpinnings for the remainder of the paper. Although methodological pluralism takes diverse forms, it suits our modelling approach to focus on two diverging properties of evidence: "sharpness" and "independence". We'll first briefly discuss method pluralism, before spending some time developing and clarifying these two evidential properties and briefly pointing to examples where these notions are plausibly in scientific play.

A "scientific method" is a strategy for doing science: how should scientists go about generating knowledge? The kind of "disunity" or "pluralism" that concerns us here denies that science can be unified by a single method[3] (see Feyarabend 1975 for a classic defence of this claim). That is, there is no one privileged strategy for generating scientific knowledge that is distinctive of it. Rather, method is context-sensitive. There is room for disagreement about which aspects of context matter, and to what extent: some emphasize the social and political,[4] others research's public import,[5] others the kinds of systems the investigation targets.[6] Regardless, that good science admits a plurality of method — that it is "disunified" — we take to be generally agreed upon.

In our model, agents adopting different strategies compete by trying to maximize different properties of evidence. We'll call one "sharpness", the other "independence". It is useful to begin with the old truth that observations do not count as evidence for or against a hypothesis *simpliciter* bodies of background theory are required to connect our hypothesis-driven expectations with an investigation's results.[7] The relationship between background theory and evidence underwrites two evidential properties.

On the one hand, evidence can be more or less *sharp*. Brandon (1997) understands how "experimental" a study is in part by the extent to which experimental results test the relevant hypothesis. The epistemic "power" of a result, or set of results, to establish or falsify a conjecture, varies. Sharpness, then, is a relationship between *results* and *hypotheses*. Dull results are ambiguous *vis-à-vis* hypotheses; sharp evidence speaks clearly and firmly. For this to occur, background theory must connect the investigation's results to the hypothesis, and alternatives must be accounted for.

Sharpness is best understood probabilistically. Evidence is sharp *vis-à-vis* a hypothesis to the extent to which the unconditional probability of the hypothesis is lower than its likelihood given the evidence in question. That is, sharp evidence has a high likelihood ratio. In Bayesian terms, "sharpness" is a measure of how much we should update our beliefs in light of the evidence. Dull evidence will raise our credence but a little, while sharp evidence has major effects.[8] It is plausible that many experimental strategies aim to maximize sharpness (consider Cleland 2002, Currie & Levy forthcoming, for instance). Controlling for confounding factors lowers ambiguity; the results exclude more possibilities. Multiple runs, controlled conditions, and other features of experiments make for powerful, convincing, *sharp* results.

[3] It is worth pointing out that this disunity thesis is weaker, and does not necessarily come hand-in-hand with the stronger metaphysical claims of Dupré (1995) and Cartwright (1999).

[4] See, for instance, Shapin & Schaffer (1985)

[5] For instance, Douglas (2000), Brown (2013).

[6] Currie & Walsh (2018), Weisberg (2013), Matthewson (2011).

[7] See, for instance, Bogen & Woodward (1988).

[8] Things are not as simple as they may seem here: on many Bayesian accounts, evidence for a proposition has diminishing returns — that is, if my confidence is already very high, evidence which in other contexts might be very telling would, on this account, be dull. There are several moves open here. One might adopt a *four-place* account of sharpness, relative to (1) the hypothesis in question, (2) the observations or data, (3) background theory, and (4) the current beliefs of the relevant agents (this could be a three-place account if background theory determines current beliefs). Alternatively, one might adopt a counterfactual account, understanding sharpness as the amount by which a rational agent's credence would increase in light of the evidence *if* their priors were suitably low. Or, one could abandon the subjectivism of Bayesianism and adopt an externalist account of evidence: thus, there is some non-agent-relative fact of the matter about sharpness. There may be subtle differences in the construal of our model based on these decisions, but as we discuss in footnote 21, our model will cohere with a wide variety of precisifications.

On the other hand, variety-of-evidence reasoning relies on *independence*. We can understand "independence" as the extent to which lines of evidence rely on varied background theories. Consider proxies of past temperature. Surface temperature fluctuations in the deep past can be detected by (among other things) boreholes, and preserved pollen grains in sediment. First, the temperature variation at different vertical positions of boreholes tracks temperature variation at the time of deposition. Second, pollen quantity tracks plant productivity, and as this is sensitive to temperature, fluctuations in pollen quantity can tell us about temperature fluctuations across time. Again, data are not evidence for free: borehole temperature must be controlled against warmth from the Earth's core, for instance. So, both proxies require background theory for evidential relevance, but — crucially — they require *different* background theory. That is, they are *independent*, and independence can sometimes carry important epistemic consequences (see Forber & Griffith 2011; Fitelson 2001; Heesen, Bright, & Zucker 2014; Stegenga 2009).[9]

Imagine that both borehole and pollen data converge on the same pattern of past temperature. Because the evidence relies on different bodies of background theory, for the world to refuse to cooperate — that is, for the convergent predictions to turn out false — distinct failures are required. If temperature estimates from both borehole and pollen data converge, but those estimations are false, separate mistakes are required for each source. Perhaps our method for pollen-gathering introduces bias; perhaps interior warming in our analysis of borehole data was faulty. In this circumstance, the convergence would be a, perhaps very unlikely, coincidence. Independence, then, is a virtue. Under the right conditions, if the measures converge on the same result, then it is likely to be the *right* result, as otherwise both must mess up but coincidentally converge. If, however, data rely on overlapping background theory, then a single mistake can lead to the failure of both. Independence is graded: some measures will have more or less overlapping justification. And, indeed, overlaps can be more or less problematic depending on how firmly established the overlapping theory is. Although independence is in principle a virtue, it is important to note its in-practice limitations. Evidence generated from different procedures can be *incongruent*: background theory is required to "translate" between evidence generated by different procedures using different language (Stegenga 2011). Further, often different evidence is used to support different aspects of a hypothesis: they merely cohere rather than converge in the sense independence requires (Currie 2018, ch. 8; also Mayo-Wilson 2011).[10]

Nancy Cartwright (2007) makes a related evidential distinction, and it is worth clarifying the difference between her approach and ours. She distinguishes between *clinchers* — a form of evidence which is very strong (deductive) but narrow in scope — and *vouchers*, broader evidence which simply adds inductive weight, but doesn't "clinch" the deal for the hypothesis in question. Clinchers are narrow because of their deductive nature: "The assumptions necessary for their successful application will have to be extremely restrictive and they can take only a very specialized type of evidence as input and special forms of conclusion as output" (12). Evidence's sharpness, or the independence of a set of evidence, doesn't entail whether we should think of that evidence as a clincher or a voucher: that turns on how restrictive the assumptions, and how special the outputs and inputs, must be for evidential relevance. However, as we'll mention in section 6.1, highly sharp evidence is likely to have the restricted scope that Cartwright identifies with clinchers.[11] We can further specify sharpness and independence using formal machinery. We don't think the formal machinery is strictly-speaking necessary for our purposes, and less technically minded readers might prefer to go straight to section 3.

[9] The term 'robustness' is sometimes used in discussion of 'independence'; we prefer the latter term, as 'robustness' is also used to discuss virtues of models and experimental setups which are not directly related to the evidence they generate.

[10] Thanks to an anonymous referee for suggesting these clarifications.

[11] Thanks to an anonymous referee for pointing us towards Cartwright's distinction.

Consider two situations. In situation 1, we are trying to establish whether hypothesis h is true. Our prior is low, say p(h)= 0.01 (perhaps h is a putative causal pathway, and there are numerous possible pathways consistent with our background knowledge). By investing resources, we can increasingly refine a method m that will generate data set D. Given background knowledge, we know that if the data set yields evidence that confirms our hypothesised pathway, that evidence will be powerful, say $p(h|e) = 0.95$. Furthermore, our background knowledge could indicate that the data set D is likely to produce evidence for one, and only one, of the competing hypotheses we are currently entertaining within set H, such that, say,such that, say,

$$p(h_i) = 0.01 \cap p(h_i|e_i) = 0.95 \cap p\left(h_{j\neq i}\middle|e_i\right) = \varepsilon \cap p(\neg \bigcup_k e_k \,|D) = 0.05, h_{i,j} \in H$$

Then we'll say method m is very sharp.

In situation 2 we will have two methods, m_1 and m_2, generating data sets D_1 and D_2. Following notation from above, the situation here will be of dull evidence:

$$\forall D, p(h_i) = 0.01 \cap p(h_i|e_i) = 0.04 \cap p\left(h_{j\neq i}\middle|e_i\right) = 0.03 \cap p(\neg \bigcup_k e_k|D) = 0.8, h_{i,j} \in H$$

However, we will add a special boost to the posterior in the case of evidential convergence:

$$p\left(h_i\middle|e_{i,D_1} \cap e_{i,D_2}\right) = 0.95 \cap p\left(h_{j\neq i}\middle|e_{i,D_1} \cap e_{i,D_2}\right) = \varepsilon$$

In such a case, we will say the two methods m_1 and m_2 are highly independent of each other. There are other ways one might formalize independence,[12] but the simulation we introduce in section 3 can represent most of these.

So, we can distinguish between how sharp evidence is, that is, its lack of ambiguity *vis-à-vis* a hypothesis, given background theory; and its independence, that is, the amount of overlap between background theory pertaining to different evidence sources. In our model, both of these properties will be represented on a landscape. The former will be the height dimension of the landscape — higher values will represent sharper evidence. The latter will be represented as distance on the landscape. The distinction also allows us to explain the two strategies adopted by agents on the landscape.

[12] There are different ways of characterizing independence (thanks to an anonymous referee for encouraging us to expand on this point). Stegenga & Menon, for instance, distinguish between probabilistic independence (which is closer to what we have in mind) and ontological independence (2017). Although we discuss independence in terms of (lack of) overlap in background theory, there are other ways to go within a Bayesian framework, and subtleties to how one formalises it with respect to different measures of evidential support (Fitelson 2001). We take independence to be a relationship between two (or more) instances of evidence (E1 and E2) and a hypothesis (H) for which they can potentially provide confirmation (and not, it is important to note, a two-way relationship between instances of evidence, or methods for generating them). However, a datum does not become evidence for (or against) a particular hypothesis without the background knowledge (K) that connects the datum to the hypothesis and allows it to perform the role of evidence: $P(E|H\cap K)>P(E|\neg H\cap K),P(E|H\cap\neg K)\approx P(E|\neg H\cap\neg K)$. Thus we get to our notion of independence: it is the extent to which the parts of the background knowledge (K_1 and K_2) that underpin the confirmation relationship between the different data and the hypothesis are independent of each other, such that a fault in one would not undermine the strength of the other; for independent sources of evidence, $P(E1\cap E2 \,|\, H\cap K1\cap K2)> P(E1\cap E2 \,|\, H\cap K1\cap\neg K2)>P(E1\cap E2 \,|\, \neg H\cap K1\cap\neg K2)$, whereas for non-independent sources the second inequality does not necessarily hold. Another way to look at independence is in terms of the hypothesis screening off the probabilistic connection between the different data (Sober 1989). For independent sources of evidence, it is the truth of the hypothesis (rather than some other aspect of the world) that makes both data turn out in a way that supports the hypothesis. On this view, for independent sources of evidence, $P(E2 \,|\, E1)> P(E2 \,|\, E1\cap H)$. As we'll discuss below, we think our approach to modelling independence (and sharpness, for that matter) handles the majority of approaches to precisifying these notions.

In our model, we will distinguish between two scientific strategies: one attempts to maximize sharpness, the other independence. The former strategy is followed by *methodological obligates*: they seek out sources that generate maximally sharp evidence. The latter is followed by *methodological omnivores*: they seek to minimize the overlap in background theory between the evidence they have.[13] Clearly, these strategies are major simplifications of actual science; the distinction is drawn for the purposes of modelling. However, we do think that some differences in scientific methodology do reflect the obligate/omnivore distinction.

First, scientists interested in uncovering the past often emphasize the need to "do science differently" in the face of a lack of experimental access to their targets and the decay of past information (see, for instance, the introduction to Diamond & Robinson 2010, as well as Turner 2007). In light of this, the extent to which one can rely on a single or a few sources of evidence dramatically diminishes: scientists instead adopt a "variety-of-reasoning" strategy which seeks to maximize their epistemic reach. Philosophical accounts of historical reasoning often emphasize the importance of independence *in light of* a lack of access to sharp evidence. Because, in such contexts, evidence is often ambiguous, biased, and degraded (in our terminology, *dull*), scientists weave together several independent evidence sources (see, in particular, Currie 2016, 2018; Wylie 2011, Chapman & Wylie 2016; Forber & Griffith 2011; Vezér 2016).

Second, ecologists are often worried about the legitimacy of their evidential practices (for a classic example, see Weiner 1997). As William Bausman has recently discussed (Bausman 2016, under review; Bausman & Halina forthcoming), "neutral theorists" complain that ecological methodologies which focus on competition models are epistemically inadequate, because they lack the crucial tests provided by null models. They argue that models without competition (neutral models) should be used to test competition models. Competition modellers, in response, point to the empirical fruitfulness of their approach. Where one side emphasizes statistical testing, the other points to the use of "natural experiments" (Diamond & Robinson 2010). So, competition theorists approach an ecosystem by positing a set of trophic interactions between populations in that ecosystem: patterns of abundance are explained in light of interactions between, for instance, predators and prey. As evidence, they cite those population-level patterns themselves, and less direct evidence from a variety of sources which suggest that, in effect, such patterns are often due to trophic interactions. In response, "neutral modellers" demand it be shown that those same patterns cannot be generated by models which do not posit trophic interactions. We think this debate is plausibly read as a demand for sharpness on the part of the neutral theorists, and a defence of variety-of-evidence reasoning from competition theorists.

Third, defenders of Evidence-Based Medicine are plausibly read as demanding sharpness, and denying the value of independence, in the context of approving medicinal treatments. On such views, the best evidence (sometimes, in effect, the only admissible evidence) for proving the effectiveness of a treatment is a randomised controlled trial and, ideally, a meta-analysis of such trials. These are contrasted with anecdotal, narrative, and lab-based mechanistic types of evidence which are considered less important. Others respond that medicine would do better to take a "total evidence view", including these other evidence sources in approving medical treatments (for general discussion, see Stagenga 2011, Solomon 2015). Here, the evidence-based-medicine folk appear to be demanding a certain sharpness, while their detractors think that independence matters too.

To show whether our distinction really captures these debates and others like it, of course, would require significantly more argument on our part. The point of this tour is to provide some preliminary reason to think that the sharpness/independence distinction has some claim to plausibility in practice. In what follows, we'll contrast two features which might make a difference to such debates. First, there is what we'll call "evidential context" — understood narrowly as concerning

[13] The obligate/omnivore distinction is adapted from Currie (2015, 2018).

the effectiveness of those strategies given the nature of the target systems involved, the available evidence, and so forth. Second, there are the beliefs and values of the epistemic community at hand. We aim for our model to explore how a community's beliefs can shape both the kind of evidence generated, and the variety and productivity of that community. An upshot of this discussion is the reminder that debates about evidence do not occur in a social vacuum: understanding why scientists approve of what they do, and how they progress, is not a simple matter of considering the appropriate strategy given an evidential context. Social factors matter — crucially — as well.

3. Modelling Method Pluralism

Here is not the place for a full defence of the use of agent-based models in institutional design, but we'll make a few observations before discussing our model. If, as we argue, method bias is a real phenomenon with serious real-world implications, this paper could form part of an agenda calling for the redesign of various scientific institutions affected by this bias (e.g. publishing and funding institutions). But how does such simulation work feed into such agendas? We'll by and large follow Roth (2002) and Alexandrova & Northcott (2009)'s discussions.

The model presented in this paper is driven not by empirical data, but by idealised representations of "reasonable assumptions" about the target domain. Therefore, it may be useful to think of the model as a formalised thought experiment. Both thought experiments and our model operate by making a complex system "concrete" (in the sense of *specific*, not in the sense of *actual*). However, unlike thought experiments, which concretise by loading a hypothetical anecdote with what are taken to be exemplary characteristics, the model concretises by assigning numerical parameters to what are taken to be key processes.

In the design of the model, the usefulness of the concretisation relies on our judgements of what is reasonable and what is important. The designer's judgement is called upon twice: in the choice of relevant processes, and in their numerical parametrisation. Due to its reliance on largely untested beliefs about what is relevant and what is reasonable, the model is not predictive. Nonetheless, in the best scenario, it can serve as a template for predictive hypotheses, once the relevant data have been gathered. In this capacity, it can also serve as a guide to data collection, prioritising some data-gathering activities over others. Further, the concretisation that model-building provides allows us to see how our ideas might interact, thus revealing important connections between them.

A useful way of approaching questions about diversity in science co-opts landscape models from evolutionary biology.[14] A standard evolutionary landscape consists of three dimensions, X, Y, and Z. X and Y form a two-dimensional grid: locations representing genotypes. A third dimension — Z — adds a topography to this grid, representing the fitness of various genotypes. Agents explore the landscape according to various rules. In evolutionary landscapes, agents are typically "hill-climbers", shifting from lower to higher locations on the grid. This is useful for representing, for instance, local fitness traps: an agent may reach local optima but, due to "valleys", be unable to reach higher ground.[15]

Philosophers and sociologists of science have reconceived such landscapes in epistemic terms. Typically (following Weisberg & Muldoon 2009, see also Grim 2009; Grim at al 2013; Alexander et al 2015; Thoma 2015) locations on the X, Y grid represent topics that a scientist might decide to pursue, while the Z axis represents the significance of a result — i.e. the local optima might represent a publication in a top journal. In this context, philosophers have asked which search strategies are more likely to locate peaks in the landscape: typically "follower' strategies, which piggyback on already explored locations, are contrasted with "maverick" strategies, which prefer unexplored areas.

[14] Wright (1932)

[15] Though see Gavrilets (2004) for criticism of this use, stemming from the pernicious simplifying use of low (two) dimensions to represent a high-dimensionality space. We note that this criticism doesn't bite as strongly for our model, since our agents are not local hill-climbers, as described below.

In our case, we're interested in a different set of questions: first, the relationship between epistemic situation and evidence-gathering strategy; second, the relationship between evidence generation and method bias. As such, our model differs from previous work both in terms of its dynamics — as we'll see, these are more complex — and in terms of construal. Where X, Y coordinates represent topics for Weisberg & Muldoon, for us they represent *methods*: particular investigative techniques. Where the Z axis previously represented significance, we'll take it to mean sharpness, such that the height of each location is the sharpness of the evidence produced by that particular method. Further, distance between X, Y coordinates in our model represents the overlap between background theory which underwrites methods — that is, independence.

A common criticism of existing epistemic landscape models is that neither height of individual points on the landscape nor distance between points on the landscape have rigorous philosophical underpinnings.[16] In our model, both of these parameters are clearer. Each parameter (height and distance) maps directly to the goals of obligates and omnivores respectively, namely sharpness (degree of belief update following evidence) and independence (degree of information overlap between two different kinds of evidence, given background knowledge).[17] As such, we conceive of each location, given by a specified (x, y) coordinate, as representing a method of data generation, and of landscape "height" (Z axis) as evidence sharpness: the higher the point, the sharper the evidence produced by the method, given background knowledge. Distance in the landscape — on the X, Y plane — represents independence:[18] the further apart two methods are, the more independent their evidence outputs will be from each other, i.e. less shared background theory goes into linking the evidence streams to a hypothesis.[19] The basic idea is that obligate and omnivore agents pursue research programs preferring sharpness or independence. By manipulating landscape topography and publishing requirements, we can examine the relationship between methodological strategy, epistemic situation, and method bias.

What do we take our model to be a model of? At minimum, a population of scientists are interested in which evidential sources and tests will lead them to the truth concerning some particular hypothesis or set of hypotheses. One group of scientists — the obligates — bet that sharp evidence is the way to go, while another — omnivores — seek out independence.[20] Our model captures a set of minimal conditions for when two different features of evidence might matter within an investigation. One way of capturing these minimal conditions is by

[16] See criticism of current lack of solid foundations in Avin (2015), ch. 2. See Avin (2015), ch. 3, for an attempt to provide such a foundation.

[17] Though note possible complications arising from the subjective nature of background belief and belief update, as noted in footnote 8, and from various interpretations of independence, as noted in footnote 12.

[18] As long as the conceptualisation of independence takes the form of a three-way relationship, or more generally an (n+1)-way relationship for a hypothesis and n instances of evidence, and as long as independence is taken to be gradual (such that overlap of background knowledge, or amount of screening, can vary along a spectrum), we can associate, for a given hypothesis (or set of mutually inconsistent hypotheses) the degree of independence with a distance metric, and use it to map various evidential sources onto a landscape. We use this property in the simulations that follow, and our results should hold for any conceptualisation of independence that allows associating a distance metric to a collection of evidence instances or sources, *vis-à-vis* a hypothesis under consideration.

[19] Given this notion of distance we also chose a bounded, non-toroidal topology for our model.

[20] We can view this situation through Bayesian eyes. There's a set of hypotheses under consideration, and a range of evidential sources which could provide confirmation of one hypothesis against others. We care about two properties of evidence in relation to the hypothesis set. One is a direct confirmation relationship that maps an evidence source and a hypothesis set to a degree of confirmation for the best-supported hypothesis in the set $f1\ (e,\{h\})\rightarrow c\ (h_{best})$. The other is a relational confirmation relationship that maps a set of evidence sources and a hypothesis set to a degree of confirmation for the best-supported hypothesis in the set $f2\ (\{e\},\{h\})\rightarrow c(h_{best})$. In our model we cash these out in terms of *sharpness* and *independence* respectively, but our simulation results should hold for any way of factoring the confirmation relationship into two complementary relationships, one that esmphasizes the direct link between an evidence source and a hypothesis, and the other that emphasizes the relations between evidence sources (all of this, of course, should take into account general- and evidence-source-specific background knowledge).

appealing to the Kuhnian notion of *normal science*. That is, our model captures circumstances where there is more-or-less agreement about what hypotheses matter, how evidence effects those hypotheses, and so on.[21]

As mentioned above, our landscape is a three-dimensional configuration space. Agents interact with the landscape, generating evidence and publishing their results. Agents are distinguished by how they do so: obligates prefer to pursue sharpness, and will publish when sufficient sharpness is reached; omnivores will prefer independence, and will publish when they collect a body of evidence that spans a sufficiently diverse background. We can represent method bias by manipulating publishing requirements.

The model is evolutionary: after set time periods, the ratio of obligates to omnivores is altered, with some agents "defecting" to the "winning" strategy, where "winning" is determined by the number of publications attributed to each strategy. We don't mean for these evolutionary dynamics to represent the development of actual scientific communities over time. Rather, the revealed evolutionary trajectories show how different publishing requirements and landscape topographies favour different research strategies.

In designing the model, we ran calibration simulations which explored ranges of reasonable parameter values to find combinations of landscape topography, agent strategies, and publication thresholds that consistently result in landscapes where both obligates and omnivores survive in dynamic equilibrium near a 50%/50% population split. We used these parameter configurations to establish "neutral" landscapes and publication thresholds that we could then vary to explore the effects of method mismatch and method bias. The values which determine, for instance, agent behaviour are therefore not arbitrary, at least in this minimal sense.

Let's highlight some model features one-by-one.[22]

Landscape values

We utilize the three dimensions of the landscape (distance along the X dimension, distance along the Y dimension, and height along the Z dimension) to represent the two distinct qualities of evidence discussed above: sharpness and independence. Our landscape is dynamic, with sharpness values allowed to vary over the course of the simulation.

The Z axis (sharpness) of a landscape consists in both a potential value or "ceiling" and an actual value. At the beginning of a simulation run, the potential value is created by adding randomly shaped bivariate Gaussians to a flat landscape. The initial actual value is then set to some fraction of the potential value. By "generating" (see below), agents can increase the exploitable landscape to above the initial actual values, but it can exceed the ceiling only under special circumstances.[23] When agents "exploit" (see below), the actual value along the Z axis decreases. When the actual value of sharpness changes, it changes for all agents in the model; there is only one unchanging "potential" landscape throughout the simulation and, at any given simulation step, only one "actual" landscape, which are shared by all agents.

In addition to sharpness — the Z value — we're also interested in independence. This is roughly the distance between two locations on the X, Y grid. Distance is an infamous source of trouble for landscape models.[24] In evolutionary models, it is plausibly read as similarity between genotypes; in epistemic models, similarity between investigations or techniques. But in what sense are genotypes or techniques

[21] More explicitly, we take a straightforward model interpretation to require (1) general community agreement about the hypothesis set under consideration $\{h\}$, (2) an *a posteriori* agreement about the admissible evidence set $\{e\}$ (this is *a posteriori* because evidential sources are initially unknown to the community members, but we require that, once a method is discovered, its results are accepted as evidence by the knowledge gatekeepers of the community), and (3) agreement within each sub-community pursuing one of f_1 or f_2 about their values (which can be cached out, for example, by demanding agreement about the background knowledge required for each evidence source).

[22] Both the model's code and an expanded explanation of the model's variables and operation can be found online at https://github.com/shaharavin/method-bias/.

[23] If two peaks overlap, the total value can exceed the ceiling.

[24] Stadler et al (2001), Avin (2015, ch. 3).

similar; and is this similarity reasonably represented in two dimensions? On our construal, distance is a measure not of similarity, but of overlap between background theories which underwrite methods. We don't doubt that there is much more to be said in working out precisely what this amounts to, but insofar as it relies on a notion of overlap rather than similarity, it is, we think, an improvement.

So, distance is conceived as a measure of independence: methods close to each other in the landscape have significant overlap in background theory, while those far apart have less overlap. When extending from distance, a two-way relationship, to an *N*-way relationship, we measure the independence of a set of methods as the area of the polygon bounding the coordinates corresponding to these methods. An important variable regarding distance is the *clustering* of peaks. An initial variable determines to what extent peaks are distributed within the landscape: high values lead to peaks clustering in the centre, such that the area of the polygon bounding the highest peaks is small relative to the size of the landscape (low independence); low values of clustering lead to a wider distribution of peaks, a larger bounding polygon, and higher independence. Unlike sharpness, the degree of independence of different methods (the X and Y coordinates of methods) remains fixed throughout the simulation; however, as sharpness varies, sometimes a method becomes so dull ("0" sharpness) that it is no longer worth pursuing, at which point it would not contribute to independence, and so the potential for generating independent evidence on the landscape changes over time even as distances remain fixed.

Agent behaviours

At the beginning of the simulation, agents are seeded on the landscape in random locations. Each turn, agents perform one of four behaviours, and an agent's strategy (omnivore or obligate) dictates how likely they are to perform each behaviour. If a behaviour is selected that the agent cannot perform that turn, they select a new behaviour with the unavailable behaviour removed. Let's examine each behaviour.

Exploration

In the model, each agent tracks their own "vision", representing the methods each scientist knows about and can deploy to generate evidence; vision is not shared between agents. At a simulation run's beginning, each agent's vision is restricted to a radius around their initial position.[25] When an agent *explores*, they shift their position to a random unexplored spot on the landscape (allowing "hopping") and update their vision. We can imagine exploration as a set of quite disparate scientific activities: literature reviews, field work, and so forth. As the simulation proceeds, agents uncover increasing amounts of the landscape. Agents never lose vision of previously explored areas, and can "see" changes to sharpness caused by other agents (see below).

Exploitation

Exploitation mines evidence from the landscape. Agents are unrestricted spatially — they can exploit any visible (to them) spot on the landscape. On exploitation an agent scans known locales and selects the best spot according to their strategy (see below). Once a spot is selected, the evidence for that position — its sharpness and location (which affects independence) — is added to the agent's evidence stack. Exploitation also changes topography to reflect both diminishing returns in evidential sharpness and loss of novelty. The actual height of the position is decreased by a fixed fraction of the ceiling. Note that reduction is shallow and surgical: the landscape doesn't reduce completely, and the surrounding landscape is more-or-less unscathed. We can imagine exploitation as scientists producing data: analysing fossils, say, or running an experiment.

Generation

Scientists do not simply run experiments; they also must design them. Palaeontologists do not only analyse fossils; they must find them, and

[25] We used a radius of 5 for vision updates, using Moore neighborhood (all grid locations within distance 5 and within the grid's bounds).

produce new methods for doing so. *Generation* is a way of modelling this activity. When an agent generates, actual landscape value increases both at their current position and at a randomly selected spot. The extra spot is to represent the indirect upshots of new techniques. The two peaks are both generated at the potential value of their respective locations. Generation is time-consuming, but agents get first dibs on accessing the fruits of their labour — after generation is complete, the agent immediately exploits.

Submission

Once an agent's evidence stack meets a certain threshold (threshold is strategy-dependent — see below), they are able to submit their evidence for publication. The submitting agent's evidence stack is cleared, and a new publication with that evidence emerges. The publication records both which agent produced it and the strategy of the agent. Publication quantity per strategy determines the ratio of strategies at the beginning of the next episode (see below).

Agent strategies

Agents adopt one of two strategies, defined by how likely they are to perform actions (their *weighting*), and by the sufficiency conditions for publication.

Omnivores are scientists who value independent evidence: they are likely to seek out new methods, techniques, and data sources, and this is reflected in their tendency to prefer exploration. Their bar for publication submission is determined by the independence of their stack of evidence. In the first experiment, they may publish once the area of the bounding polygon of their evidence stack is greater than or equal to 1/5 of the landscape area (the X, Y dimensions). They are twice as likely to explore as they are to generate, exploit, or submit. This represents "variety-of-evidence" reasoning. When an omnivore exploits, they select locations which are maximally distant from the locations in their evidence stack (so long as that location has at least some non-zero value in terms of sharpness).

Obligates are scientists who value sharpness: they are less likely to look for different kinds of evidence, but instead prefer to focus on developing a promising area. Their bar for publication is determined by the *total sharpness* of their stack of evidence. In the first experiment, this is when the sum of the evidence sharpness is greater than three times the landscape's ceiling. They generate and exploit twice as often as they submit, and explore only half as often as they submit. This represents the kind of reasoning we see in much experimental science: careful, controlled studies are the focus. Obligates select the highest visible peak when exploiting.

Table 1 Agent strategies and associated behaviour likelihoods and publication criteria

	Explore Prob.	**Generate Prob.**	**Exploit Prob.**	**Submit Prob.**	**Publish (in 1st and 2nd experiment)**
Obligate	1 / 11	4 / 11	4 / 11	2 / 11	Total evidence sharpness > 3x the ceiling
Omnivore	2 / 5	1 / 5	1 / 5	1 / 5	Polygon area of evidence > 1/5 landscape area

Simulation Runs

At the beginning of a simulation, various initial parameters are set, including agent number, landscape size, and so forth. Agents are seeded, and each performs one action per turn (unless they have generated, at which point no action is taken for six turns). At the end of an episode (which is a set number of turns), the ratio of omnivores to obligates is updated to reflect the relative success of each strategy in publishing papers. We take a turn to represent three months of research: a very coarse estimate of a field-season or an experimental run.

Figure 1 shows a series of snapshots from a sample simulation run. Variations in colour indicate landscape values, with the blue end of the spectrum representing low values, and red high. Between snapshots, exploitation leads to reduction in local sharpness (less red, more blue), while generation leads to increases in sharpness. Locations obligates exploit are marked with black dots and cluster around the highest peaks; locations omnivores exploit are marked with white dots and tend to occur on the periphery of the landscape.

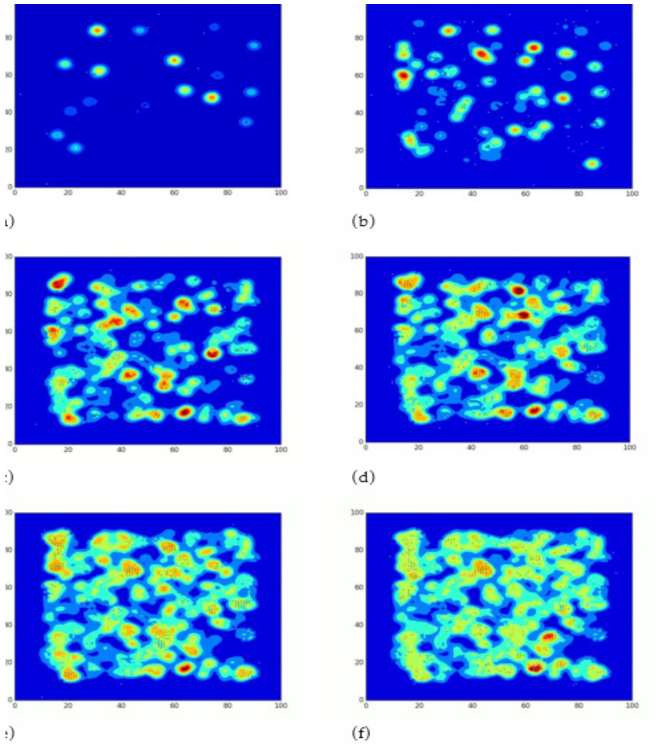

Figure 1. Snapshots of a sample simulation run. X- and Y-axes map to landscape coordinates, and colours represent height (from dark blue, which is lowest, to dark red, which is highest). The locations where agents exploit are marked on the landscape as well, with black dots for obligate exploitation and white dots for omnivore exploitation. The series shows the landscape dynamics following generation and exploitation.

4. Experiments: Topography & Strategy

The literature on methodological pluralism implies that caring about independence or sharpness — that is, adopting omnivore or obligate strategies — is in part a response to epistemic situation. We can understand an epistemic situation as concerning, first, an evidential context — that is, the kinds of evidence sources which are available — and, second, a broader set of social influences. Our landscape is intended to represent the former (we examine the latter by shifting publication requirements in the next section). Obligates should flourish when there are rich seams of evidence to exploit (clustered, tall peaks), while omnivores should do well when the epistemic landscape is more diffuse and sparse. We ran two experiments to establish that, in our model, the success of a strategy is sensitive to landscape topography, which reasonably represents evidential context.

Experiment 1: Abundance

What's the relationship between total landscape *abundance* (that is, the sum of heights of all landscape coordinates) and whether the landscape favours obligate or omnivore strategies? A circumstance with bountiful, sharp evidence should encourage scientists to focus on sharpness. When evidence is duller, a strategy of focusing on variety-of-evidence reasoning is more appropriate. And indeed, in our model, increasing abundance leads to higher obligate favouring, and lower abundance leads to higher omnivore favouring (see fig 2).

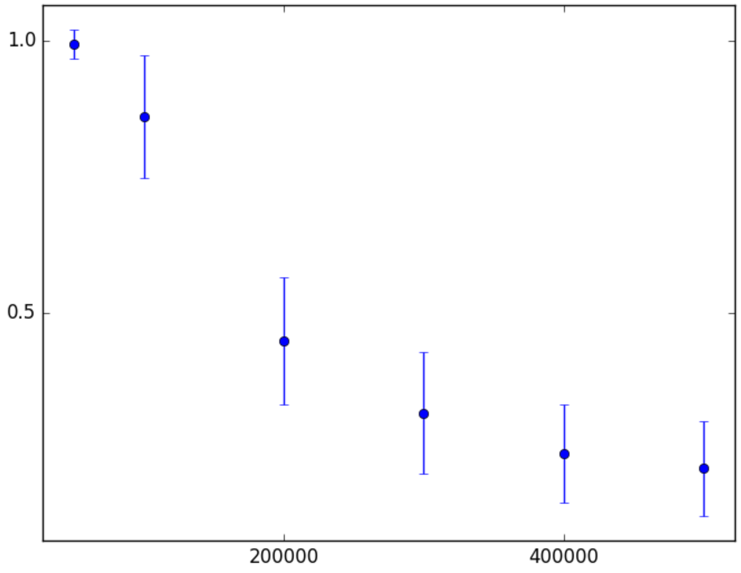

Figure 2: Plot of omnivore ratio relative to the total population at simulation end (after 150 steps), as a function of total landscape abundance, averaged over 50 runs (error bars show one standard deviation).

Figure 2 represents fifty simulation runs per value of total landscape abundance. The vertical axis represents the ratio of omnivores to the total population; the horizontal axis represents landscape abundance. There is a clear pattern: landscapes of low abundance are dominated by omnivores; at higher values, obligates are favoured. These results are reason to think that our model is behaving as it should: sharpness-rich circumstances encourage obligate strategies.

Experiment 2: Clustering

In a further experiment on the relationship between topography and strategy, we altered the spread of value in the landscape. We in effect kept total abundance and landscape ceiling steady,[26] but varied the distribution of value across the landscapes. We predicted that landscapes with concentrated value at or near the centre would favour obligates, while landscapes with value spread across the landscape would favour omnivores. By manipulating landscape clustering while holding abundance fixed, we thus examined the relationship between how distributed evidence is, and the success of strategies (fig 3)

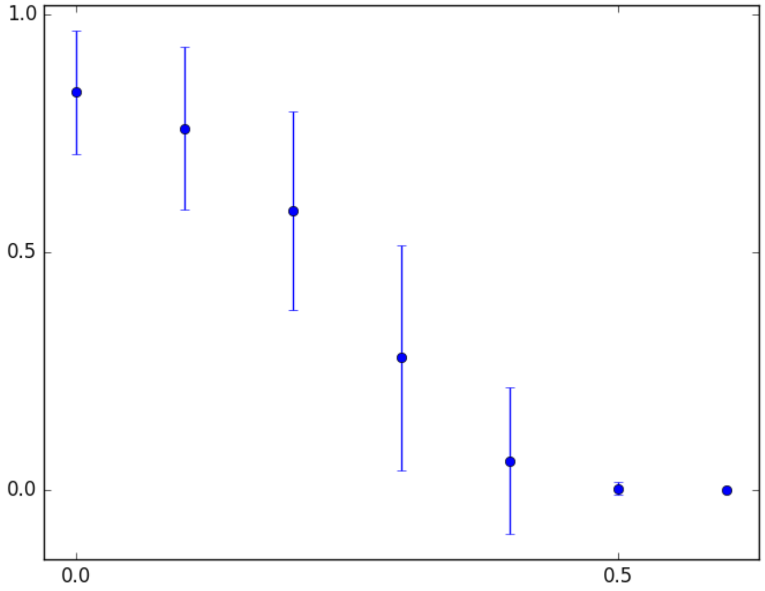

Figure 3: Omnivore ratio as a function of peak clustering.

[26] There is some variation due to peak overlaps.

Again, in the figure, the vertical axis represents omnivore ratio to the total population at the end of the simulation. The horizontal axis represents clustering, where low values place no restriction on initial peak placement, and higher values increasingly restrict peaks to the centre of the landscape. We ran fifty simulations for each value of the clustering parameter shown in the figure. Sure enough, increasing clustering favoured obligates, and lowering it favoured omnivores.

5. Experiments: Bias

The first two experiments established the relationship between strategy success and landscape topography. In effect, this shows that our model behaves as it ought: landscapes with abundant, sharp, clustered evidence favour obligates; unabundant, dull, dispersed landscapes favour omnivores. We construe this as representing the kinds of differences between sciences discussed in section 2. Where some scientists seek "rich veins" of sharp evidence, others adopt variety-of-evidence reasoning to mitigate the ambiguity of their data. Having established that different strategies were favoured in different topographies, we now introduce publishing bias to see how it affects favouring. We can introduce bias by shifting the sufficiency for publication in one strategy but not another. Recall that epistemic situations include both evidential contexts — represented by the landscape — and broader social aspects. We take publication bias to coarsely represent these broader aspects — and indeed, our models are intended to show how crucial these are.

We should disambiguate two ways in which publishers might be "unbiased". On the one hand, a "laissez-faire" bias takes an unbiased publisher to go, as it were, with the evidential context. They will not attempt to interfere with the "natural" path of things. On the other hand, a "balanced" notion of bias understands unbiased publishers as striving for an even split between strategies. They will sometimes work *against* the evidential context to ensure that as close as possible to equal proportions are present. These differing conceptions are unproblematic for our purposes: "bias" is simply the comparative ease of publishing for the various strategies (irrelevant of topography). Whether bias is egregious, indifferent, or positive is not built into our model — this depends on context and the goals at hand.

Experiment 3: Method Bias

We know that certain landscapes favour certain strategies: clustered, abundant landscapes favour obligates; dispersed, sparse landscapes favour omnivores. Can method bias, here in its guise of publishing standards, work against or mitigate the strategies favoured by the landscape? To answer this, we conducted a series of experiments where the publishing requirements for one strategy were increased while those for the other remained steady.

In the experiment below, we track omnivore ratio against an omnivore publishing bias in a landscape which highly favours obligates (high clustering and abundance). Low values on the horizontal axis represent low standards for omnivore publishing — roughly, the lower the value, the easier it is for omnivores to publish.

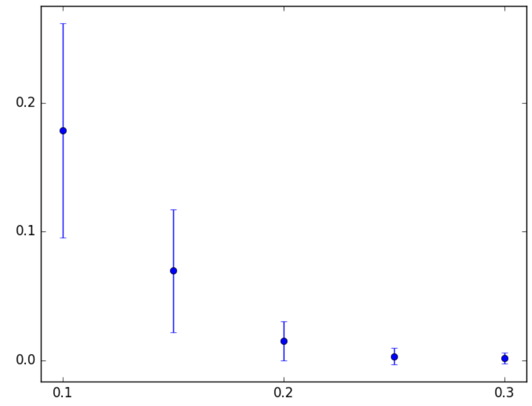

Figure 4: Omnivore ratio as a function of omnivore publication threshold on an obligate-favouring landscape.

In an epistemic landscape so favouring of obligates, things must be made *very* easy to keep omnivores in the population. Even at our highest publishing bias, under ¼ of the population are omnivores. However, the bias does keep them in the population — where, under usual conditions, topography would exclude omnivores, diversity can be maintained using publishing bias.

Further (and unsurprisingly), publishing bias can also affect both the amount of sharpness and independence extracted from a landscape.

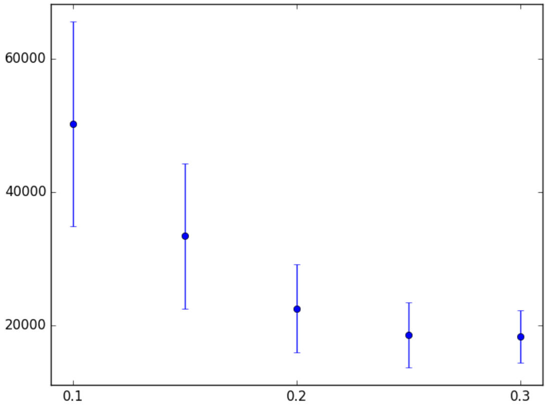

Figure 5: Total independence summed across all publications as a function of omnivore publication threshold on an obligate-favouring landscape.

This graph charts omnivore publishing bias on the horizontal axis (as before) but this time tracks the total amount of evidential independence on the vertical axis. The results are remarkably similar to those previously shown (see figure 4), and this makes sense: as publishing bias keeps more omnivores in a population, and those omnivores focus on producing highly independent evidence, we should expect the total population to produce more independence of evidence.

The lesson here is that publishing bias can affect evidence generation in two ways. Firstly, it can maintain methodologically mismatched strategies in a population which would otherwise be eliminated. Secondly — and in virtue of this first feature — it can increase the amount of evidence quality that is mismatched to the evidential context: bias can increase the independence extracted from abundant landscapes, or sharpness from diffuse landscapes.

In the above experiments, we have mitigated the context-matching advantages of the obligate by making things easier for the omnivores. In the next study, we leave the omnivores' publication thresholds as they are, but make it trickier for the obligates.

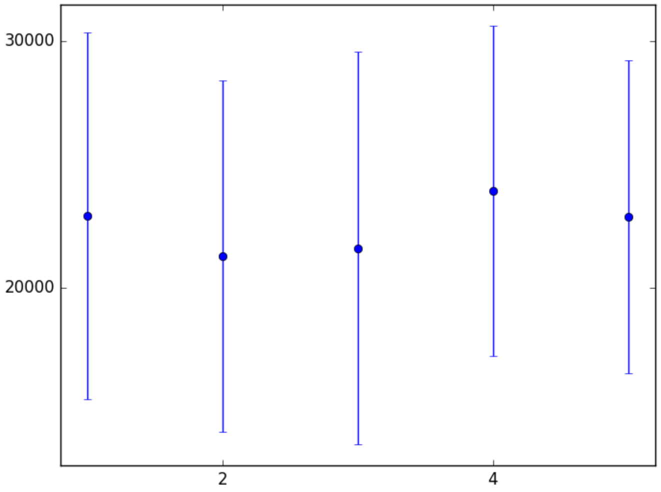

Figure 6: Total independence as a function of obligate publication threshold on an obligate-favouring landscape.

Figure 6 plots total independence (vertical axis) against obligate-bias (horizontal axis) in an obligate-favouring landscape. We've seen above that total independence in this kind of landscape is a good proxy for both epistemic and strategic diversity. We see that there is no discernible pattern — making things easier or harder for the context-matching obligates doesn't seem to make a systematic difference to the amount of independence produced. A tempting lesson to draw here is that publishing bias produces more diversity when *favouring the mismatched* than when *disfavouring the matched*. As mentioned above, such conclusions are not prescriptive, both because the model is agnostic with respect to the valence of bias, and because it is too idealised for direct application to science policy.

Nonetheless, we offer here a template for a causal hypothesis that could be tested: lowering evidential requirements for the method which is less effective in the evidential context (e.g. a variety-of-evidence paper in a traditional experimental-sharpness-oriented discipline) would have more effect than hardening the requirements for a favoured method. Though conceptual and methodological hurdles abound, such a hypothesis could conceivably be put to the test.

Efficiency cost

A further question: what costs are there in overall productivity, and in terms of the evidence quality pursued by the *context-matching* strategy when publishing bias is introduced?

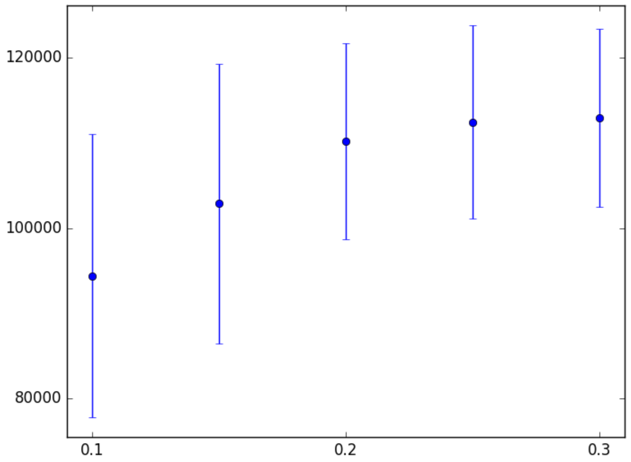

Figure 7: Total sharpness across all publications as a function of omnivore publication threshold on an obligate-favouring landscape.

The above tracks total sharpness (vertical axis) against omnivore bias (horizontal axis), again in an obligate-favouring landscape. Undoubtedly, the amount of sharpness extracted from a landscape decreases at more extreme omnivore publishing biases (and thus omnivore ratio, and total independence). Interestingly, the average amount still increases — although not enormously, once error bars are taken into account — once the bias is lowered to the point where, in the previous studies, omnivores left the population. That is, the previous experiments showed us that at 0.2 there are next to no omnivores in the population, and yet the trend in quality continues its upwards trajectory to 0.25. This suggests that biasing against the mismatched methodology can have positive effects on the context-matching methodology — presumably because it decreases the number of "lucky" omnivores who happen to hang on beyond expectations. Bias can, then, operate in both directions: it can increase how *focused* an epistemic community is, by conspiring *with* the epistemic landscape; or it can increase *diversity*, by working *against* the landscape.

Is the trade-off a fair one? That is, when we encourage the mismatched methodology, do we discourage the other in equal amounts? We think not (and this is, presumably, because omnivores still produce some sharpness, and obligates some independence). To draw a comparison, we ran a series of experiments which looked at a neutral landscape, an omnivore-favouring landscape, and an obligate-favouring landscape. The values in the neutral landscape were used as a metric against which the others were valued. The table below tracks the results:

	Neutral bias	**Omni bias (for)**	**Obli bias (for)**
Neutral	1,1	0.68,1.38	1.72,0.47
Omni-favouring	0.31,1.69	0.31,2.04	0.46,1.54
Obli-favouring	11.37,0.17	9.08,0.38	11.93,0.16

The first value represents total sharpness, the second total independence. The values are indexed to the neutral values. So, for instance, the score of 11.93 in the obligate-favouring, obligate bias landscape, in the bottom right corner, represents a nearly 12-fold increase in the total amount of sharpness extracted from the (much more abundantly sharp) landscape. These results suggest that there is not a fair trade-off in such cases — in fact, the cost paid in the advantaged value (the property sought by the context-matched methodology) is *less than* the amount gained in the disadvantaged value. In an obligate-favouring landscape, we pay a cost of 20% reduction in total sharpness (from 11.37 with no bias to 9.08 with omnivore-favouring bias), for a 124% increase in the amount of independence generated (from 0.17 to 0.38). A similar effect is seen on the omnivore-favouring landscape. There, a 9% reduction in total independence (from 1.69 to 1.54) is paid to generate a 48% increase in sharpness (from 0.31 to 0.46).

The lesson, we take it, is that although there are costs associated with publishing bias, these are neither simple nor fair. In our model, at least, the introduction of bias against the grain does not have equal costs for the advantaged value — potentially, then, overall efficiency (taking into account both sharpness and independence) of the community has increased. Of course, in our model we have no non-arbitrary way of *combining* independence and sharpness into a single measure, and so 'overall efficiency' is not meaningful in that context.[27] Regardless, the discrepancy between increases in one value and decreases in the other are suggestive. If this trade-off can be confirmed empirically, it could support an argument for methodological pluralism in epistemic communities — there may be diminishing returns for a particular strategy, and much to be gained by having at least a small number of researchers pursuing another.

6. Discussion

We've aimed to provide a systematic way of thinking about method pluralism, to argue that methods might not match evidential context, and to explore how bias can both undermine and aid us in increasing the efficiency and diversity of epistemic communities. Before concluding, we want to make two points. First, we'll consider the potential downsides of method bias in various contexts. Second, we'll suggest that consideration of method bias should lead us to think that community standards about what good science looks like are a necessary component of philosophical explanations of scientific evidence.

Egregious Method Bias

Our model does not tell you when method bias is egregious or advantageous; it only generates the effects publishing bias might have

[27] It might be argued that a lack of a unified measure is problematic, but we're not so sure. First, it's unclear to us whether the kinds of evidence which scientists often bring together are non-arbitrarily combined in practice. Second, our focus is not on the evidential value of a landscape *per se*, but on how different publishing practices might affect the make-up of populations. Thanks to Remco Heesen for pushing us on this point.

vis-à-vis methodological pluralism. Whether we want to emphasize one kind of evidence, or want a diversity of strategies, or whatever, depends crucially on what we want to do with the scientific evidence. A circumstance which concerns us is when method bias acts to diminish exactly the kind of evidence we want. For example, an emphasis on highly focused — sharp — experimental evidence could lead us to misunderstand how our results will play out in the complex, interdependent world beyond the lab. Where the costs of getting things wrong matter, and they often do, getting it right involves understanding the limitations of the evidence we are able to generate.

Egregious method bias and mismatch are particularly problematic in circumstances where scientific results are, as it were, in the public eye: when they matter for public policy, for instance. Preferring a kind of evidence which is inappropriate to context could result in mismanagement, and misunderstanding the stability, accuracy, or trustworthiness of scientific claims. Our technological prowess is plausibly outpacing our scientific understanding, and it becomes increasingly difficult to understand the impact that interventions (intentional or not) might have on complex, large-scale systems. Considering method mismatch and method bias is crucial for debates about the validity of scientific studies that attempt to ascertain the effects of climate change, the effectiveness of medicinal treatments, the safety of new AI technologies, and so forth (Avin et al. 2018). In these contexts, we should ask whether the kinds of evidence we demand and want are appropriate to both the task at hand, and the kinds of questions and systems we're interested in knowing about. Insofar as scientific results guide policy, preferring one sort of evidence or approach when another is more appropriate could be disastrous.

Often, the more-or-less unambiguous results of methodological obligates are seen as the gold standard of scientific success, but in some contexts this method is inappropriate. And often these places are just where such risky gaps in our understanding occur. As Nancy Cartwright makes particularly vivid, firmly understood knowledge in highly controlled experimental settings often collapses once it leaves the safe confines of the lab (1993, 1999, 2007). The hunt for "clinchers" leads to stable results, but these are very limited in their application. Transporting knowledge from laboratory settings into the wild often involves dramatic switches in epistemic context — plausibly from a context encouraging an obligate strategy to one favouring an omnivore strategy. That is, applying our hard-won knowledge of how things behave in controlled settings — where obligate strategies often pay dividends — to the world outside requires variety-of-evidence reasoning, as the sheer increased complexity and heterogeneity dulls the evidence in the new context. Under these conditions, method bias could lead us to both ignore routes to better discoveries and to misjudge the importance and reliability of the information we do have.

Knowledge Generation & Community Standards

A further consequence of our discussion of method plurality, bias, and mismatch concerns what a philosophical account of scientific knowledge should be like. Philosophers have often approached scientific evidence *narrowly*: the philosophical task *vis-à-vis* scientific evidence requires understanding confirmation — that is, explaining the relationship between observations and hypotheses. It strikes us that consideration of method bias puts pressure on such narrow conceptions.

One (admittedly caricatured) illustration of a narrow conception appeals to the distinction between contexts of justification and discovery. Originally coined by Reichenbach (1938, although his distinction was quite nuanced — see Schickore 2014), the distinction was used to carve out a place for philosophical analysis *vis-à-vis* science. Roughly speaking, *discovery* is the processes by which scientists conceive of, and come to, scientific theories, as well as the business of generating evidence. *Justification* involves understanding the connection between evidence and theories. One way of defending a narrow conception of science is to say the latter, rather than the former, is the proper target of analysis (as Popper argued in *The Logic of Scientific Discovery*, 1934). The logical and abstract questions of justification are the philosopher's playground, while the messy, human side can be relegated to the dustbin of "discovery". As such, when philosophers consider science, we should enquire after the connection between theories, evidence, and the world — understanding those issues provides the philosophical essence of science.

Philosophers have applied pressure on these views for a long time. The narrow view misses important aspects of scientific practice necessary for understanding its very success, progress, and stability. Such arguments take multiple forms. One set argues that justification itself has non-epistemic properties, often due to inductive risk (Douglas 2000). Another set argues that justification is found in places usually associated with discovery: specifically, the social organization of science plays an important role in preserving and supporting its stability (Longino 1990). A third set argues that scientific goods, the outputs and aims of their investigations, are not limited to well-supported theories — scientists are also interested in building storehouses of data (Hacking 1983), provisioning understanding (Potochnik 2017), and so on. What unifies these critiques is the claim that narrow conceptions are just too narrow.

Our critique is complementary: as evidence comes in a range of flavours, and (crucially) in different contexts some flavours perform better than others, combining them into a single relation between hypotheses and theories obscures the different work they do.[28] And indeed, this becomes particularly problematic in light of method bias. If different communities have different ideas about what good science is like, and use these to guide how the community develops, understanding different evidential properties and scientific strategies is necessary for understanding that evidence. That is, the relationships between evidential context and community standards are an essential part of a philosophy of scientific knowledge generation: they are part of the context of justification.

To see this, recall our Bayesian sketches of independence and sharpness. Read narrowly, pieces of Bayesian evidence just are observations which grant reason to update our subjective priors pertaining to relevant hypotheses. Powerful formalisms are bought to the fore to demonstrate how various aspects of evidence can be incorporated into this probabilistic picture. And indeed, this is often an enlightening and rich way of proceeding.[29]

Although Bayesian machinery can precisify what we mean by independence and sharpness, it doesn't follow from this that the machinery captures what matters about sharpness or independence. First, it does not tell us under what epistemic circumstances sharpness or independence ought to be favoured. That is, although the Bayesian can combine the two measures, and accommodate them, she cannot link them to evidential context. To do this, we would need to either represent something like our epistemic landscape or, in a less abstract mood, characterize the actual conditions scientists are working under and the actual aims they have. A Bayesian precisification might explain why, for instance, variety-of-evidence reasoning can be epistemically powerful, but it cannot explain why scientists might need to "do things differently" in different contexts. That is, it cannot explain why one group of scientists adopting an obligate strategy is the right thing to do, while another group adopting an omnivore strategy is the right thing to do. Epistemic situations, then, are a crucial part of the context of justification.

Second, such a precisification cannot explain community standards — method biases — that shape how a scientific community behaves. That is, in addition to missing epistemic context, they also cannot accommodate epistemic situations. And our modelling demonstrates how important such biases might be for the epistemic success of those communities: by maintaining a diversity of strategies, for example, or for targeting particular evidential properties. Understanding

[28] We think analogous arguments can be found. Toulmin's position that scientific reasoning should focus on warrants rather than logical relationships (1958) and Norton's defence of a material theory of induction (2003) both can be read as making something like this argument.

[29] Wallach (2016), for instance, has no problem fitting diverse archaeological evidence into a Bayesian framework.

scientific evidence involves understanding how a community's conception of good method shapes what work in fact gets done.

As with other objections to narrow conceptions, we do not claim that such work is without value, nor that it fails to achieve many of its aims. Rather, the point is that the perspective is itself limited — and even limited in terms of its narrow concerns of evidence and confirmation — as such, we take consideration of method pluralism to further a more pluralistic, increasingly local approach to science from philosophers themselves.

7. Conclusion

Our aim in this paper has been, first, to introduce the notion of method bias as it arises from a recognition of method pluralism; second, to explore these notions using a series of simulations; and, third, to sketch some consequences for philosophical understanding of scientific knowledge. The extent to which lessons from the second aim can be exported into actual practice depends crucially on the extent to which the model's simplifying assumptions matter for features of the real-world systems — scientific communities — that we're interested in. Models like ours must be contextualized to have direct empirical consequences. Regardless, the experiments motivate a series of further, potentially testable, claims about scientific communities. Insofar as the epistemic landscape will favour certain kinds of studies, publishing bias can affect this by increasing efficiency, diversity, or both. Moreover, it is plausible that positive discrimination works better than negative discrimination — make it easier for the little guy, not harder for the big guy. The trade-offs faced are not simple; sometimes at least the introduction of method bias of the sort discussed here will have greater gains than losses. Such considerations, we think, motivate a richer, more local understanding of the nature of scientific evidence and confirmation.

Bibliography

Alexander, J. M., Himmelreich, J., & Thompson, C. (2015). Epistemic landscapes, optimal search, and the division of cognitive labor. *Philosophy of Science*, 82(3), 424–453.

Alexandrova, A. & Northcott, R. (2009). Progress in economics: Lessons from the spectrum auctions. In: Kincaid, H. & Ross, D. (Eds.), *The Oxford Handbook of Philosophy of Economics*. Oxford Handbooks Online, Ch. 11, pp. 306–337.

Avin, S. (2015). Breaking the grant cycle: On the rational allocation of public resources to scientific research projects. PhD thesis, University of Cambridge, Cambridge, UK.

Avin, S., Wintle, B.C., Weitzdörfer, J., Ó hÉigeartaigh, S.S., Sutherland, W.J., & Rees, M.J. (2018). Classifying global catastrophic risks. *Futures*. doi:10.1016/j.futures.2018.02.001.

Bausman, W.C. (under review). Modeling: Neutral, null, and baseline.

Bausman, W.C. (2016). Neutral theory, biased world. PhD thesis, University of Minnesota, Minneapolis, MN.

Bausman, W.C. & Halina, M. (forthcoming) Not null enough: Non-statistical null hypotheses in community ecology and comparative psychology. Biology & Philosophy.

Bogen, J. & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, 97(3), 303–352.

Brandon, R.N. (1997). Does biology have laws? The experimental evidence. *Philosophy of Science*, 64, S444–S457.

Brown, M.J. (2013). Values in science beyond underdetermination and inductive risk. *Philosophy of Science*, 80(5), 829–839.

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties*, 2(1), 11–20.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press.

Chapman, R. & Wylie, A. (2016). *Evidential Reasoning in Archaeology*. Bloomsbury Publishing.

Cleland, C.E. (2002), Methodological and epistemic differences between historical and experimental science. *Philosophy of Science*, 69(3), 474–496.

Currie, A. (2018). *Rock, Bone, and Ruin: An Optimist's Guide to the Historical Sciences*. The MIT Press.

Currie, A. (2016). Hot-blooded gluttons: Dependency, coherence, and method in the historical sciences. *The British Journal for the Philosophy of Science*, 68(4), 929–952.

Currie, A. (2015). Marsupial lions and methodological omnivory: Function, success and reconstruction in paleobiology. *Biology & Philosophy*, 30(2), 187–209.

Currie, A. & Levy, A. (forthcoming). Why experiments matter. Inquiry.

Currie, A. & Walsh, K. (2018). Newton on islandworld: Ontic-driven explanations of scientific method. *Perspectives on Science*, 26(1), 119–156.

Diamond, J. & Robinson, J.A. (Eds.). (2010). *Natural Experiments of History*. Harvard University Press.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, 67(4), 559–579.

Dupré, J. (1995). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Harvard University Press.

Feyarabend, P. (1975). *Against Method*. New Left Books.

Fitelson, B. (2001). A Bayesian account of independent evidence with applications. *Philosophy of Science*, 68(S3), S123–S140.

Forber, P. & Griffith, E. (2011). Historical reconstruction: Gaining epistemic access to the deep past. *Philosophy and Theory in Biology*, 3, 1–19.

Gavrilets, S. (2004). *Fitness Landscapes and the Origin of Species*. Princeton University Press.

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.

Grim, P., Singer, D.J., Fisher, S., Bramson, A., Berger, W.J., Reade, C., Flocken, C., & Sales, A. (2013). Scientific networks on data landscapes: Question difficulty, epistemic success, and convergence. *Episteme*, 10(4), 441–464.

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science* (Vol. 5, No. 1). Cambridge University Press.

Heesen, R., Bright, L.K., & Zucker, A. (2014). Vindicating methodological triangulation. *Synthese*, 1–15.

Leonelli, S. (2016). *Data-centric Biology: A Philosophical Study*. University of Chicago Press.

Longino, H.E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Matthewson, J. (2011). Trade-offs in model-building: A more target-oriented approach. *Studies in History and Philosophy of Science Part A*, 42(2), 324–333.

Mayo-Wilson, C. (2011). The problem of piecemeal induction. *Philosophy of Science*, 78(5), 864–874.

Norton, J.D. (2003). A material theory of induction. *Philosophy of Science*, 70(4), 647–670.

Popper, K. (2002 [1934/1959]). *The Logic of Scientific Discovery*. Routledge. First published in German in 1934; first English translation in 1959.

Potochnik, A. (2017). *Idealization and the Aims of Science*. University of Chicago Press.

Reichenbach, H. (1938). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. The University of Chicago Press.

Roth, A.E. (2002). The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica* 70(4), 1341–1378.

Schickore, J. (2014). Scientific discovery. *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), Edward N. Zalta (Ed.), URL = <https://plato.stanford.edu/archives/spr2014/entries/scientific-discovery/>.

Shapin, S. & Schaffer, S. (1985). *Leviathan and the Air-Pump*. Princeton University Press.

Solomon, M. (2015). *Making Medical Knowledge*. Oxford University Press.

Stadler, B.M., Stadler, P.F., Wagner, G.P., & Fontana, W. (2001). The topology of the possible: Formal spaces underlying patterns of evolutionary change. *Journal of Theoretical Biology* 213(2), 241–274.

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 42(4), 497–507.

Stegenga, J. (2009). Robustness, discordance, and relevance. *Philosophy of Science*, 76(5), 650–661.

Stegenga, J. & Menon, T. (2017). Robustness and independent evidence. *Philosophy of Science*, 84(3), 414–435.

Thoma, J. (2015). The epistemic division of labor revisited. *Philosophy of Science* 82(3), 454–472.

Toulmin, S.E. (1958). *The Uses of Argument*. Cambridge University Press.

Turner, D. (2009). Beyond detective work: Empirical testing in paleontology. In: *The Paleobiological Revolution: Essays on the Growth of Modern Paleontology*, Sepkoski, D. & Ruse, M. (Eds.). University of Chicago Press, Ch. 10, pp. 201–214.

Turner, D. (2007). Making prehistory: Historical science and the scientific realism debate. Cambridge University Press: Cambridge.

Vezér, M.A. (2016). Variety-of-evidence reasoning about the distant past. *European Journal for Philosophy of Science*, 7(2), 1–9.

Wallach, E. (2016). Bayesian representation of a prolonged archaeological debate. *Synthese*, 195(1), 401–431.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

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

Wylie, A. (2011). Critical distance: stabilising evidential claims in archaeology. In: Dawid, P., Twining, W., & Vasilaki, M. (Eds.), *Evidence, Inference and Enquiry*. Oxford University Press/British Academy, Ch. 14, pp. 371–394.