Existential Risk, Creativity & Well-Adapted Science

Penultimate Version, forthcoming in Studies in the History & Philosophy of Science.

Adrian Currie

Abstract

Existential risks, particularly those arising from emerging technologies, are a complex, obstinate challenge for scientific study. This should motivate studying how the relevant scientific communities might be made more amenable to studying those kinds of targets. I offer an account of scientific creativity suitable for thinking about scientific communities, and provide reasons for thinking contemporary science doesn't incentivise creativity in that sense. However, a successful science of existential risk will be creative in my sense. If we want to make progress on those questions, then, we should consider how to shift scientific incentives to encourage creativity. The analysis also has lessons for philosophical approaches to understanding the social structure of science. I introduce the notion of a 'well-adapted' science: one in which the incentive structure is tailored to the epistemic situation at hand.

Acknowledgements

I'm grateful to Shahar Avin, Marta Halina, Thi Nguyen and two anonymous referees for very helpful comments on earlier drafts. This paper was presented at the *Risk & the Culture of Science* workshop in Cambridge, the comments there helped formed the paper, as did discussion with my colleagues at the Centre for the Study of Existential Risk. 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 author(s) and do not necessarily reflect the views of Templeton World Charity Foundation.

1. Introduction

I'm worried that contemporary science is insufficiently creative to handle some of the more extreme, if improbable, risks from emerging technology. To capture my worry, we'll need to consider the *social epistemology* of science[1]. Where traditionally philosophers took the locus of knowledge to be the individual—what should *I* believe given my sensory evidence, say—social epistemologists recognise that epistemic agents are fundamentally social agents. This is crucial for understanding science: the capacity for scientific learning, publishing and dissemination depends on interconnected networks, databases, and institutions. Whether or not we think the fundamental locus of knowledge, where it lives, is at the individual or group-level, in attempting to understand or explain the epistemology of science, the isolated scientist peering into a microscope is an impoverished starting point. To understand science, we should look to the group.

The *economic approach* is increasingly popular in the social epistemology of science (Patha & David 1994, Muldoon 2013)[2]. Individual scientists are taken to be incentive governed agents—credit-maximizers—and scientific communities are modelled on this basis, using simple equations or more complex simulations. This perspective allows us to test the robustness of common platitudes about good science: that there ought to be a division of labour[3], or that information should flow freely[4], for example. My approach differs but is complementary. As opposed to using formal models, I'll use a thick description of what I call a scientific endeavour's *epistemic situation*: roughly the conditions of knowledge generation and the challenges facing it[5]. In doing

[1] I take foundational work in the social epistemology of science to include Longino (2002), Solomon (2001), Kitcher (1990, 1993) and Strevens (2003).

[2] Take my use of 'economic' with a pinch of salt: I think this is a useful, if imperfect term for a set of more-or-less formal approaches to understanding science. Although much of it does take explicit cue from economics, some does not: landscape models, for instance, are adapted from evolutionary biology.

[3] For instance Weisberg & Muldoon (2009), Thoma (2015), Pöyhönen (2016), Alexander et al (2015).

[4] For instance Zollman (2010, 2012).

[5] For discussion see Leonelli (2016, 7.3) and Currie (2018 chapter 1).

so, I'll draw on insights from the economic approach and—I hope—inform it as well. I'll suggest that the capacity for simple models to inform, guide, and understand scientific practice is amplified by contextualization in the manner I illustrate. Specifically, the lessons of such models ought to be put in contact with the epistemic situation at hand, and in some contexts fine-grained, specific detail might make a difference. For the social epistemology of science, then, local details matter. Further, philosophers are increasingly concerned with the role of non-epistemic values in science: their role in setting evidential standards, and in distributing epistemic resources (Douglas 2009, Kitcher 2001, 2011). Bringing these two thoughts together, I'll introduce a notion of science being *well-adapted*. A research program is well-adapted when the standards, incentives and expectations governing investigations are geared towards overcoming the challenges of the relevant epistemic situation. As we'll see, the notion of a well-adapted science helps integrate work both on scientific values and the economic approach.

I'll make my argument by analysing a case which is both urgent and, I'll suggest, challenges science's adaptedness: the study of human species-level threats, or *existential risk*. Paradigm existential risks have a similar profile: they are more-or-less unprecedented, large-scale, complex, and improbable. Understanding such risks (let alone knowing how to mitigate them!) requires creative multi-disciplinary work. Communicating such risks—given their Hollywood-blockbuster-potential—requires delicacy. This all makes for a tricky epistemic situation, particularly considering that contemporary science is geared towards the conservative rather than the creative.

My intention, then, is to both inform philosophical reflection on the social epistemology of science, and to plea for a better-adapted science of existential risk. I'll begin with an account of scientific creativity which philosophers following the economic approach will find familiar. I'll adapt recent work on creativity in human development to distinguish two modes of problem-solving, and suggest this can be adapted to understand the creativity of scientific communities. In

short, this consists in *cold searches*, where close locations within a solution-space are methodically examined, and *hot searches*, involving leaps across solution-space. For my purposes, the former counts as conservative and the latter as creative. With this in place, I'll argue that the social organization of science encourages cold searches. I'll then turn to existential risk, describing the epistemic situation at hand, and arguing that such a situation demands hot searches. I'll close with a discussion of my two themes. First, science's incentive structures should be well-adapted to local conditions, suggesting that the economic approach is most informative when contextualized to an epistemic situation. Second, I'll consider the challenges facing a science of existential risk: what does a well-adapted science of low-probability, high-impact events look like? This last discussion is more a promissory note then a set of concrete policy suggestions: my aim is to clearly identify the challenges faced, and thus set and motivate a research program into how those challenges might be met or mitigated.

2. Scientific Creativity

'Creativity' is polysemous[6], and I won't attempt an exhaustive account of its guises. Rather, I'll provide a definition which is (1) grounded in current science, (2) lends itself to social epistemology, and (3) illuminates contemporary philosophy of science[7]. Drawing on recent work in developmental psychology, itself taking cue from artificial intelligence, I'll characterize scientific creativity in terms of how a solution space is explored. I'll distinguish between 'hot' and 'cold' searches. These terms are inspired from thermodynamics, metaphorically referring to the kinetic energy of a molecule (a scientist) or a collection of molecules (a research community). A

[6] See, for instance, Gaut (2010), Paul & Kaufman (2014).

[7] In philosophical accounts dealing with creativity more generally, mine has most affinity with Margaret Boden's concept of 'exploratory creativity' (2004), and perhaps clashes most directly with Berys Gaut's arguments that creativity ought to be an agential property (2010)—see the introduction to this issue (Currie, this issue) for further discussion of the relationship between creativity in philosophy and the concept developed here.

hot molecule, with plenty of stored energy, will move through a space in erratic bounds. A colder molecule will move more slowly. How the metaphor plays out should become obvious.

Gopnik et al (2017) explore problem solving across different life history stages using a simple experimental paradigm. The underlying thought is that *H. sapiens*' distinctive long childhood might serve to facilitate an individual's adaptation to the various environments it might be confronted with. Because different cognitive powers are better for learning about an environment, and for behaving optimally within it, an early 'exploratory' phase could be followed by a less flexible, but better adapted 'exploitative' phase. As they say,

> …there may be a developmental trade-off between cognitive abilities that allow organisms to learn the structure of a new physical or social environment, abilities that are characteristic of children, and the more adult abilities that allow skilled action in a familiar environment. (7892)

Gopnik et al draw a connection between this cognitive tradeoff and one from artificial intelligence – between *exploration* and *exploitation.*

> Reinforcement learning algorithms make an important distinction between periods of exploration, in which the system gathers information about potential actions and outcomes, and exploitation, in which information gathering is replaced by taking the actions most likely to maximize reward. (7893).

As an individual (artificial or otherwise) adapts to her environment, a tradeoff must be struck between learning the lay of the land—exploring the space—and making use of that knowledge—coming to efficient solutions[8]. Gopnik et al provide experimental evidence that although younger children are typically outperformed when it comes to efficiency, in circumstances where solutions are based on unusual patterns of reasoning, children outdo adults.

[8] See Gopnik et al (2004), Tenenbaum et al (2011). It's worth pointing out that exploration of necessity involves some exploitation in many contexts: we should think of these strategies not in terms of whether exploitation or exploration occurs, but the size of the problem-space considered for exploitation.

One way of fleshing out hot and cold searches appeals to Bayesian priors. Bayesian agents have priors which determine their credence in propositions. Take the numerical sequence 1, 3, 5, 7… My immediate guess as to the next numeral is '9'. This is because I have certain expectations—priors—which lead me to favor certain kinds of patterns over others: even though nothing about the sequence *logically demands* that '9' be the correct answer (Norton, draft chapter 2)[9]. My predicting 9 is an example of a *cold* search: given some solution space, I'll probe a relatively small area of solutions. A *hot search* would put less weight on priors – I will be more likely to try something out, even if I haven't tried it before, or even if it has failed in the past. Our priors serve to set expectations across a space of possible solutions to a problem. Cold-searching agents will methodically exhaust their local solution-space; hot-searching agents will 'jump' about the landscape. From this notion we can define an agent's creativity as follows:

> **An agent's creativity is proportional to the likelihood of that agent attempting a distant solution, where a solution's distance is indexed to the agent's priors.**

So, creative agents are more likely to attempt solutions at greater distances, while conservative agents will exhaust their local space: the former will be more adventurous. Understanding science, however, requires a grip on community-level processes. It may be that cold-searching individuals nonetheless amount to a creative population. If, for instance, they cluster in widely spaced groups, then despite the relative conservativeness of each agent, wide areas of solution space might be explored. It may also be possible that hot-searching agents don't lead to creativity at the group level. If, for instance, information about previous searches are lost, large jumps in solution space might not lead to the accumulation of information at the group-level. So, we should distinguish between creativity at the individual-level and the group-level:

[9] As Norton points out, even if we restrict the sequence to following a single, simple rule, underdetermination remains: the sequences could be the odd numbers, or the odd primes starting from one, or the decimal expansion of 359/2645.

A population of agents' creativity is proportional to the likelihood that those agents will explore a wide solution-space.

As noted, being constituted of creative agents is only one way that a population might be creative. Cold-searching agents with widely dispersed initial priors could also lead to a creative population, as could agents using different search algorithms (I'll illustrate further examples in my discussion of the economic approach below). Further, although I've cashed this discussion out in terms of solution-spaces, we could also think of it in terms of evidential sources (see Currie & Avin forthcoming): instead of creativity informing solutions to a problem, we can think of it as a way of generating evidence pertaining to some hypothesis.

It is important to contrast being 'creative' in my sense from being 'exhaustive'. An exhaustive search will attempt to cover every solution in the space, while a creative search will pick out distant solutions. Both creative and conservative populations may achieve exhaustive searches: the latter by systematically attempting options in a small space before slowly expanding; the former by trying wide-ranging solutions and eventually filling in the gaps. In principle, both may be exhaustive: but the search-patterns by which this is achieved would be different.

Creativity in my sense is often—but not necessarily—connected to risk. Insofar as hot searches are more costly, the strategy can involve more risk to the community or individual adopting that strategy. However, this connection is sensitive to epistemic situation. As we'll see in the next section, if we have very little idea of a landscape's topography—if we've as much chance of being located near good solutions as we are bad solutions—then wide-ranging searches could be as risky (or less risky!) than more conservative ones. Again, the risk involved, and the cost, of hot or cold searching depends crucially on the local details. This is why 'thick descriptions' of epistemic situations are necessary.

So, I've discussed a notion of creativity grounded in work in AI and developmental psychology, which lends itself to thinking about communities. Again, the account is limited. It

doesn't lend itself to understanding what we might call *ingenuity*: some creative individuals have well-trained priors about what tricks and solutions might work in rather outlandish scenarios. There's a difference between a creative search and a chaotic search, and my account is not obviously sensitive to this difference (although considering how agents update their priors might help us here)[10]. It also doesn't make room for the creativity involved in cold-searches. However, the account is suitable for our purposes here. As we'll see in the next section, it is creativity in this sense which contemporary science is ill-equipped to promote, and it is this kind of creativity which scientific study of existential risk requires. I'll next provide a short illustration before connecting my account to the economic approach.

For over a century, dinosaur systematics was founded on a central division between the Ornithischia ('bird hipped') and the Saurischia ('lizard hipped') dinosaurs: the two groups were taken to go their own evolutionary ways sometime in the mid-late Triassic. This division didn't simply matter for museum displays and documentaries, but shaped questions about the evolution and radiation of dinosaurs in the late Triassic and early Jurassic, as well as the taxonomic allegiances of fossil taxa from near the base of the dinosaur tree. Dinosaurs from the Triassic were categorized either into Ornithischia or Saurischia depending on diagnostic characters related to their hip morphology.

In 2017, Baron et al published a phylogenetic analysis which challenged this received wisdom. In short, by undertaking a wide character re-analysis they generated a phylogenetic tree that drew the basic phylogenetic division in a very different way. Langer et al's critical response failed to re-establish the old order. As Baron et al say in their discussion of that response,

[10] To put it in Gaut's (2010, 2013) terms, it doesn't include creative 'flair'.

> [the study] results in recovery of the 'traditional' topology, although with less resolution and very weak support; their result is statistically indistinguishable from the possibility that our topology provides a better explanation of the data (Baron et al 2017b, E4)

The upshot of all this is not that we should maintain the old order, nor embrace the new: rather, how the base of the dinosaur tree looks is up for grabs. This is an example of a shift from a colder to a warmer temperature science, from conservatism to creativity. Why? Previously, the space of pursuable hypotheses about dinosaur phylogeny, taxonomic membership, and dispersal was constrained to within the traditional picture. Now, it is not: and this opens the door to a much wider set of analyses, hypotheses and interpretations. Where both individual priors and community norms once constrained searches to within the Ornithischian/Saurischian phylogeny, these have been relaxed: leading to hotter searches, and thus a more creative science of early dinosaurs.

This is just one way in which the creativity (or potential creativity) of a science might increase. There, the undermining of a hypothesis opened up previously uninhabited solution space. However, that is not the only possible route: as we'll see, the social structures of science are likely a source of conservatism, and thus changing these could have similar effects. Now, to link this discussion to the economic approach.

As mentioned in the introduction, the economic approach coopts some of the tools and assumptions of economics (as well as evolutionary biology) to think about scientific communities. Tools such as analytic models and simulations are used and, often at least, scientists are treated as credit-maximizers. I'll quickly sketch two popular approaches and relate them to my account of creativity.

First, some philosophers have adapted 'bandit models' to examine scientific diversity and communication. In these models, agents pick between two possible options (each representing an arm of a 'bandit' gambling machine), where each option has a fixed probability of a fixed pay-off. Typically, one arm is 'correct' insofar as it has a higher payoff, a higher probability of payoff, or both. However, agents do not know which arm is better: rather, they have credences which

are based upon previous attempts. The model can be used to represent the trade-off between exploration and exploitation discussed above. Upon getting a good result, at what point should an agent simply focus on the lever she considers 'the best'? More 'creative' agents will take longer to settle on a lever, while less creative—conservative agents—will stick to the lever which gives them the best rewards more quickly. The price of creativity is potentially lower efficiency, while the price of conservativeness is increased possibility of getting stuck on a crappy lever. Kevin Zollman (2010) adds network dynamics to the model, in order to test whether open information is always beneficial in science: is it a good idea for each agent to be aware of each other's previous attempts? This illustrates how a population might be creative even if the agents within it are not. By limiting an agent's 'vision'—which other agents' attempts can be seen—the population as a whole becomes more likely to cover more solution space. Of course, in such models 'solution space' is small, involving only two options. Landscape models provide a wider perspective.

Epistemic landscape models consist of a three dimensional space. The X and Y dimensions form a grid. In Weisberg & Muldoon's (2009) original formulation, X-Y coordinates designate a research approach to some topic. Say, Gopnik's lab is interested in problem-solving in child development, and one location on the X-Y grid could represent the experimental paradigm they adopt. The Z axis is a series of values which add a topography to the landscape. This axis represents *significance,* peaks representing important findings. Agents are randomly placed on the landscape and, in effect, explore it attempting to find the peaks. Here, instead of having priors as in bandit models, agent-behavior is determined by algorithmic instructions. "Followers" will prefer already-explored paths, while "mavericks" will prefer unexplored paths. These models are most often used to understand division-of-labor within science: is a population with some proportion of mavericks and followers better than a homogenous one? "Better" is, as in bandit models, understood roughly in terms of the trade-off between exploration and exploitation. A population with too many followers might find itself trapped on a local optima, while maverick

strategies are likely costly (Thoma 2015, Muldoon 2017). Again, we can generate population-level creativity without creative agents. In most landscape models, agents' movement is restricted to their local neighborhood, and so mavericks cannot 'jump' to new locations. Maverick behavior, however, means that they—and thus the population overall—are likely to explore more space.

Although most uses of landscape models focus on locating peaks, Weisberg & Muldoon note that we can also be interested in maximizing the amount of explored space as well. In some cases (I'll suggest one below) we might be more interested in exploring the space of research than finding the 'significance peaks'. This is particularly the case in more rugged landscapes (that is, landscapes with many peaks). Recently, Michaels Strevens (2013) and Weisberg (2013) have modelled various features which lead to 'herding' in epistemic communities—which roughly corresponds to cold searching—and have optimistic things to say about potential interventions which encourage hotter science. I take the forthcoming discussion to be complementary.

3. Why Science is Conservative

I've defined scientific creativity in terms of hot searches, and shown how this discussion fits with recent work in the formal social epistemology of science. That work—necessarily, and not necessarily problematically—abstracts from the actual conditions of scientific investigations. This can limit what such models can achieve on their own: they can perhaps test some claims made about science (that, for instance, information-sharing is always good for scientific productivity) and can perhaps motivate more applied questions (that, for instance, a certain amount of population-level creativity can matter). However, how to bring about outcomes, and whether current scientific structures reflect creativity or not, is likely beyond the scope of such models[11]. Here, I will provide an informal argument to the effect that science generates conservative

[11] See Harnagel (this issue) for a different take on both the purpose of such models and how to improve them.

populations. That is, cold searches are promoted at the expense of hot searches. This will matter crucially in the next section when I argue that in some contexts, the study of existential risk in particular, creative scientific populations are necessary for progress. Note that I won't be here distinguishing between more local and incidental—contingent—features of scientific communities, and those which are more general and are perhaps likely to arise in any knowledge-generating practice. This distinction is often critical for understanding science's historical development, and understanding how we might shape current science, but is unnecessary for my project at this stage. My discussion expands upon Kyle Stanford's argument to a similar effect (2015). Stanford doesn't explicitly use my account of creativity, but his discussion is readily adapted[12].

And so, Stanford emphasizes contemporary science's conservativeness. In essence, he acknowledges that science is productive, but argues that this productivity isn't directed towards creativity. In my parlance, scientific incentives encourage cold rather than hot searches. Stanford's approach is based on a comparison between science prior to the 19th Century and how it is now, drawing on work in the History and Philosophy of Science such as Rudwick (1982) and Shapin (2008).

> …the professionalization of science in the middle decades of the nineteenth century, the shift to state support of academic science through peer-reviewed proposals for particular research projects following World War II, and the ongoing acceleration and expansion of so-called 'Big Science' have served to reduce not only the incentives but also the freedom scientists have to pursue research that challenges existing theoretical orthodoxy or seeks to develop fundamental theoretical innovation (2015, 7)

[12] Stanford is responding to criticisms of his (2010) arguments for scientific anti-realism on the basis of unconceived alternatives (Godfrey-Smith 2008, Forber 2008). The connection between his discussion and mine is science's conservatism: if science encourages cold searches, then it is unlikely to discover alternative theories or hypotheses.

These features lead to scientific research focusing on a relatively small part of possible research space. I'll summarize and expand upon Stanford's points, before tying this picture of contemporary scientific communities back to my account of creativity. As encapsulated in the above quote, Stanford focuses on three contrasts: the professionalization of science, the introduction of peer review, and the emergence of big science.

(1) Professionalization led to scientists becoming dependent for their ongoing work on the approval of their peers. Peer-approval determines what work is interesting, legitimate or significant—and this tends towards consensus forming regarding those matters. That is, the community becomes relatively homogenous regarding research programs, approaches and perspectives. Although the professionalization of science meant that science was open to input from a more diverse range of individuals (those from lower social classes for instance), in order to play the game one has to buy into the theoretical underpinnings which the players commit to.

(2) The advent of competitive, peer-reviewed funding after the second world war amped up conservative inertia. This gave peers more influence over what scientific work was done, thus further empowering consensus on the legitimacy or otherwise of scientific questions, approaches, and so forth. Additionally, these funding sources were often highly centralized and small in number. This encouraged less diversity in what got funded, and encouraged the funding of large projects. Further, getting funding typically requires explicit, pre-decided goals for research with likely epistemic dividends. This discourages open-ended, exploratory research (Currie 2018 chapter 12) and makes innovative, risky proposals unlikely to be green-lighted.

(3) The advent of 'big science' led to a further centralization and stratification of scientific communities. Producing data for enormous data-bases requires standardization (Leonelli 2016). It also more-or-less necessitates increasingly hierarchical structures in labs wherein scholars are unable to direct research until they are deep into their careers. Further, this leads to conservative approaches from principle investigators themselves:

> [big labs] motivates further intellectual conservatism on the part of advisors and mentors themselves, as a PI who elects to pursue a genuinely revolutionary, transformative, or theoretically iconoclastic research program is more likely to provoke skepticism from a granting agencies' program managers or review committees must now be willing to risk not only her own scientific fortunes but also those of the small army of less well-situated scientific workers whose careers presently depend upon her own. (Stanford 2015, 16).

A useful way of capturing these forces for inertia, implied by the above quote, returns us to the economic approach. Recall that by this approach we should consider scientists as credit-maximizing agents. That is, they act not to maximize their own (or the community's) knowledge-gains, but their own credit-gains. Insofar as scientific credit encourages conservatism, so scientists are led to conservative research agendas. Foster et al (2015) summarize the overall point:

> Scientists "take a position" by pursuing *particular* research problems selected from the space of all those possible. These concrete actions are guided by the interplay between scientists' positions in the field and their *habitus*: acquired systems of taste, dispositions, and expectations. At stake are recognition by fellow scientists, other currencies for which recognition can be traded, and an improved position in the field. (Foster et al, their italics, 2015, 876).

1-3 above provide features of scientific communities which make hot searches risky: they're unlikely to be funded, likely to be treated with suspicion by peers, and unlikely to be supported by large laboratories. This makes such approaches unattractive. Further, the centralization and increased hierarchy in science likely amplifies the 'Matthew Effect'. Roughly, by this effect those scientists already in eminent positions will in virtue of that accumulate still further credit[13]:

> … eminent scientist get disproportionately greater credit for their contributions to science while relatively unknown scientists tend to get disproportionately little credit for comparable contributions (Mertin 1968, 57).

This clustering of credit potentially undermines the capacity of scientists or labs which are younger, or lesser known, or who are underprivileged (in the third world for instance)—all potentially diversity increasing and thus hot-searching—to get noticed, funded, cited, built upon and so on (see De Cruz forthcoming for a discussion of such effects in philosophy). It also further puts control of scientific output to a small group. Add to this that science is a crowded marketplace: given the small number of available positions, when scientists decide which labs to join, which jobs to apply for or take, which directions to specialize in, which funding to apply for, and so forth, they are making risky bets about which directions will be successful. In a recent editorial, *Nature* reported on a survey of over five thousand early-career scientists. Of these, three quarters intended to pursue careers in science, even though "… only three or four in every hundred PhD students in the United Kingdom will have a permanent staff position at a university. It's only a little better in the United States" (2017, 429). In such an environment, it is unlikely that individuals will make things harder for themselves by attempting revolutionary hot-searches.

A further force for inertia harkens again back to Thomas Kuhn's emphasis on institutionalized learning. The process of becoming a scientist—transitioning from undergraduate, to graduate, to post-doc, and so forth—doesn't simply involve learning the skills, techniques and theories relevant to that discipline. It also involves taking on a tacit set of expectations about what research questions, approaches, and answers are legitimate and interesting, what Foster et al above called 'habitus'. This process of institutionalization likely promotes scientific productivity: after all, it allows easier communication, and grants the community a common epistemic purpose. However, it also makes it less likely that scientists will be able to overcome their tacit expectations and think revolutionary thoughts.

In addition to these more-or-less tacit expectations within a discipline, in many cases explicit standards for publication become norms. The most obvious of these is the use of p-values to set lower bars for publication across many statistical sciences, and rules against publishing negative

results (see, for instance, Benjamin et al 2017 for criticism regarding p-values). At least two upshots are relevant here: first, of the research that is done, only some will see the light of day; second, scientists will direct their research efforts towards questions which are more likely to provide results deemed significant by those standards.

Disciplinary boundaries also promote inertia. Insofar as certain disciplinary techniques and research questions constrain researchers to particular parts of research-space, without inter-disciplinary work much total space will remain unoccupied. And despite claims to the contrary, interdisciplinarity is often discourages (Bromham, Dinnage & Xua 2016): in addition to the difficulty involved in integrating work from different backgrounds, various gate-keeping processes make its success unlikely[14]. For instance, typically a discipline's most well-regarded journals are in the business of publishing papers concerning the core business of that discipline. This means that interdisciplinary work is published in less-well regarded journals, thus making them less high-profile, and thus less advantageous for career progress and attracting funding. Generally speaking, as the fundamental institutional unit of universities are by-and-large departments of particular disciplines, work on 'core' areas of those disciplines are likely to be encouraged. This all adds up to the extra effort required to integrate with people of other disciplines actually hurting one's career.

These forces for inertia are further reflected in, and reinforced by, the behaviors of scientists themselves. Scientists are often suspicious of non-mainstream investigations (or simply investigations outside of their own specialization), and often engage in informal gate-keeping behaviors. Being marked out as a maverick, or an odd-ball vis-à-vis one's scientific endeavors, serves to further isolate potentially revolutionary scientists from the community, thus decreasing their chances of being invited to conferences, being published, being funded, and so on. Such behaviors create what Huw Price has called *reputation traps*: even casual, open-minded

[14] See, for instance, papers in Mäki et al (2017)

consideration of radical scientific ideas can undermine the good name of a once respectable scientist (Price 2016). A final potential source of inertia comes from the dynamics of lab formation. Cailin O'Connor (this issue) presents formal modeling results to demonstrate that under some conditions a kind of group-selection at the lab level can drive conservativeness. In particular, if successful strategies underwriting cold-searches are more 'heritable' between labs than those underwriting hot-searches (which I think is *prima-facie* likely), then cold-searching is likely to propagate.

This all adds up to a community which pools or herds: one specialized in cold searching. Given the forces for conformity, institutional, behavioral and tacit, and given the high-risk bets involved in building scientific careers, we should expect scientists to 'play it safe'—that is, choose research paths which are likely to be respected in the community, more likely to provide epistemic dividends, and so forth. That is to say, we should expect modern science to not be creative—to encourage cold searches[15]. This in itself isn't necessarily problematic. Indeed, there are likely to be circumstances where cold searches are just what we want. However, if there are circumstances where a more creative—revolutionary—science is what we need, then modern science is ill-equipped to provide it. In the next section, I'll argue that such circumstances exist.

4. The Epistemic Situation of Existential Risk

I've thus far provided an account of scientific creativity as well as reason to think that, most of the time at least, scientific incentive structures do not encourage creativity. In my view, nothing negative follows directly from this. It is only when an epistemic situation demands a creative approach that conservative incentives are problematic: for all I've said thus far cold-

[15] This is a very general argument, and there will be local exceptions—some rather glaring. Further work needs to be done to establish to what extent efforts akin to the Nobel Prizes, say, promote riskier scientific strategies. I'm no expert on any of the sciences that get Nobels, nor on the prizes themselves, but it would be interesting to consider what effects such high-impact 'heroic' awards in fact have on scientific communities (see, for instance, Wagner et al 2015).

searching might be the best strategy for the majority of cases. In this section, then, I want to provide a case study where, I'll argue, creativity is demanded: existential risk. I won't claim that study of existential risk is unique or distinctive (far from it!) rather, I take it to be a relatively clear example of the kind of epistemic situation which demands a creative science (I'll note caveats as we go). Moreover, as an emerging discipline, characterizing it at this stage could encourage reflection on how we should conceive of that work and how it ought to best be practiced and shaped. Keep the main point in mind: I've articulated a notion of creativity linked to hot searching, argued that science doesn't encourage hot searching, and will now provide an example of an epistemic situation in which hot searching is called for.

At base, an *existential risk* (X-risk) is a threat to some thing's existence. I take a personal existential risk when I cross the road, and our species takes one when it amasses nuclear weapons[16]. Where many risks—catastrophic risks for instance—are understood in terms of scale (perhaps measured in terms of lives lost, or financial cost), existential risks are indexed to the set of things under that risk. Typically, the study of existential risk focuses on a narrow band of these risks, at the upper-end of the bell curve where we meet either human extinction (a species-level threat) or the loss of crucial aspects of civilization (a culture-level threat)[17]. Although the sources of many existential risks are not anthropogenic: extra-terrestrial impacts, supervolcanic eruptions, etc…, The focus of X-risk studies are typically risks from emerging technologies such as artificial intelligence, advanced genetic engineering technologies, and synthetic biology[18]. At

[16] This is an idiosyncratic definition: there is not, so far as I can tell, an agreed-upon account of existential risk. I prefer to distinguish existential from catastrophic risks in terms of whether the risk is indexed to the risk's subject (in this case, the existence of that subject), or to a scale (say, minor to catastrophic). Bostrom (2002, 2013) defines X-risk as one "… that threatens the premature extinction of Earth-originating intelligent life or the drastic destruction of its potential for desirable future development" (2013, 15, see Torres 2017 for expansion). I consider this a feature rather than a bug: human-level extinction risks are a motley bunch, and in different contexts different definitions may be more or less useful. I don't see any pressing reason to insist on a single orthodox definition of the domain of X-risk, so long as there is clarity within particular discussions.

[17] See Bostrom & Cirkovic (2008), Baum & Barrett (2015), Torres (2017).

[18] Although distinguishing between human-caused risks and 'natural' risks is sometimes useful, it is important to see that such distinctions have a shelf-life. The scale of 'natural' risks depends in part on what measures we have taken to understand and mitigate them.

base, our technological capacities are outrunning our capacity to understand, control or predict the consequences of employing those capacities, and as we'll see this creates a distinctive and difficult epistemic situation.

In this section then, I aim to sketch the *epistemic situation* faced by those studying X-risk. An epistemic situation consists in (1) the challenges facing knowledge generation and (2) the resources available in generating knowledge. Different disciplines and studies face different epistemic situations. Experimental biologists can conduct repeated, fine-grained experimental studies of, say, the developmental systems of fruitflies; whereas scientists testing the effects of pharmaceutical treatments rely on random controlled trials. Presumably part of the reason for the latter is the ethical unsuitability of invasive lab-studies on human subjects. Our epistemic resources and challenges are set by the nature of the systems we're studying, as well as the social, technological and ethical terrain they bump up against. This provides the kind of thick description which, I think, facilitates the contextualization of work from the economic approach.

Here, I'll focus on the challenges facing X-risk, before briefly discussing the kind of science which might meet those challenges. I'll conclude that the epistemic situation faced by X-risk demands a creative science—in part characterized by hot searches. Therein lies the conundrum at this paper's centre: the social organization of science discourages hot searches, but a science of X-risk demands them. Note that not all X-risks share the features I'll list, but I think there is sufficient overlap to be able to talk sensibly about there being a typical epistemic situation facing scientists interested in paradigm X-risks.

It is worth noting that there are non-epistemic grounds for scientists interested in X-risk to move outside of the usual thinking within their more specialized sub-disciplines. First, consider the importance of highlighting safety concerns. Raising red flags about the potential dangers of new technology is an extremely tricky business. Given the highly competitive nature of funding, and the risky bets scientists take in selecting research directions, pointing out potential risks,

especially existential ones, require individual scientists putting out their necks. If a new technology does get a whiff of the illegitimate, new researchers and funding can flee quickly (see, for instance, my discussion of geoengineering, Currie forthcoming). As such, the same forces which drive epistemic conservatism in science can also dampen the capacities of scientists working within those fields to raise and study safety concerns. Second, the global nature of both X-risk and the potential benefits of the emerging technologies which raise them likely demand that scientific and technological progress be geared towards the needs of the many, not the few. After all, as X-risks are risks for everyone, the potential benefits of technology which might raise their probability shouldn't be narrowly distributed (particularly to a privileged elite). Above I've focused on how science is epistemically conservative, but the features I've mentioned might also contribute to conservativism regarding the groups research represents the interests of. Restrictions on minority groups and those from the global south likely limit the capacity of such crucial sciences to be just. Moreover, lack of representation from those quarters likely themselves restrict scientific productivity (O'Connor & Bruner 2017).

4.1 Uniqueness

In order to build a theory or model of some phenomenon, it is *prima-facie* plausible that we require multiple examples of it. A unique, unprecedented target, then, presents an epistemic challenge: there is insufficient data to have an empirically-grounded model of the phenomena (Tucker 1998). A pertinent difference between some natural X-risks and those with anthropogenic sources is the events' uniqueness. Asteroids, volcanic activity, and so forth, leave geological signals: we can detect patterns of their occurrence, reconstruct their climatic and biological effects, and generally use the past as a guide to the present. Moreover, our species having already survived approximately one hundred thousand years without a natural event knocking us out makes it defeasibly plausible that we're safe for the next (say) hundred years

from the kind of extinction risks we faced in the past. However, man-made risks are a different ballgame:

> … our species is introducing entirely new kinds of existential risk – threats we have no track record of surviving. Our longevity as a species therefore offers no strong prior grounds for confident optimism (Bostrom 2013, 17).

Unique, unprecedented events (or possible events), then, present an epistemic challenge due to both a lack of evidence and an inability to infer from previous behavior, the result being that uncertainty about risk is likely to dominate risk assessments (Brostrom 2013).

A science of existential risk, then, must adopt techniques and strategies which mitigate a lack of evidence available to construct theories and models of the relevant phenomena.

4.2 'Wild' Systems

The systems involved in X-risk scenarios are often unfriendly to systematic scientific understanding. They are what Kirsten Walsh and I have called relatively *wild systems* (Currie & Walsh 2018). A 'wild' system is characterized as being (compared to competing systems) high in both 'interference' and 'noise'. The former concerns the interdependence of the system's parts and their effects: it is difficult to determine the causal powers of particular components in systems of high interference. The latter concerns our capacity to isolate a system: it is difficult to predict the behaviour of a target system which is open to erratic shocks from without. Wild systems—those high in interference and noise—are difficult to study because we cannot isolate and examine their components separately, and their behaviour is often irregular due to exogenous effects.

Human-extinction level threats often involve interactions between highly complex, interdependent systems. Consider extreme solar flares (Isobi 2016) The occurrence of such an event, in addition to killing the roughly half a million people airborne at any one time, would

knock out all satellites and temporarily remove the ozone layer. The effects on global trade, transport, health, politics and communication would undoubtedly be catastrophic—but how catastrophic, and how would the various knock-on effects operate? Answering such questions involves understanding not simply the inner working of particular, complex systems, but also how those behaviours would change, and themselves be changed, by their interdependencies with other systems. Both noise and interference will be high under such conditions. It's important to note that X-risks are not necessarily the outcomes of single cataclysmic events, but in many scenarios emerge from cascades of tragedy which, in combination, add up to civilization collapse or even human extinction (Karieva & Caranza forthcoming, Liu, Lauta & Mass forthcoming).

A science of X-risk, then, must adopt strategies to mitigate the noisy, high-inference nature of the systems they investigate.

4.3 2nd-Order Uncertainty

Considering uniqueness, I pointed out that uncertainty will likely dominate risk calculus pertaining to X-risks. But that is only one aspect of our ignorance: another concerns which possible events should be on our radar in the first place, and which research questions will be fruitful: to draw on the metaphor of an epistemic landscape, we are ignorant of the landscape's topography—whether there are few peaks, or a more rugged landscape—and of its dimensions: we don't know what the possible sources of X-risk are. In other words, where in 4.1 we focused on known unknowns—that is, our uncertainty regarding the likelihood of some risk—an additional and crucial aspect of our epistemic situation concerns unknown unknowns.

The space of X-risk concerns is already broad: from worries about astronomical events like asteroids and solar-flares, to politics (regarding nuclear capacities, say) to more abstract theoretical worries such as those arising from Fermi's paradox (Miller & Felton 2017). We lack systematic ways of tackling the space of X-risks (although see Avin et al forthcoming).

A science of X-risk, then, should be exploratory: ideally, systematic means of identifying possible sources of risks should be sought.

4.4 The Public Eye

In addition to challenges emerging from the nature of existential risk itself, a crucial part of the epistemic situation at hand concerns interactions between X-risk and the public.

X-risk naturally lends itself to the splashy: human extinction, the dangers of emerging technology, and so forth, make excellent fodder for science fiction and journalism alike. This brings challenges. A science of existential risk—particularly early on—will get a lot of things wrong. And indeed given the low probability of many of the events concerned, it will sometimes be hard to tell when it gets things right. This brings with it two conflicting issues. On the one hand, the public or policy-makers might take the science too seriously, and act rashly in light of that. But on the other hand, repeated potential 'failures' could lead to a loss of faith in the science.

Further, features of human psychology potentially make X-risk tricky to study insofar as any science needs at least some proportion of positive public regard. Jacob Weiner (2016) has argued that existential and other catastrophic risks face a *tragedy of the uncommons.* In these circumstances, the rareness of an event makes it likely to be misunderstood, mismanaged or neglected. Wiener suggests that, in contrast to typical situations, when facing tragedies of the uncommons experts are more likely to want regulative steps than laypeople. This is because, first, the rarity and unfamiliarity of the events make them 'unavailable' to our minds and imaginations. Second, the scale of the events likely leads to 'mass-numbing': a psychological effect where an individual's concern for some costs actually decreases as the cost increases. The effect is possibly because "… respondents feel overwhelmed and doubt that their contribution can really make a difference" (72), or because we respond more to named and known individuals

than to faceless masses. Third, our legal and other regulatory institutions are likely to be ineffective in the face of catastrophes, as they will likely break down in those scenarios, thus undermining their motivational power.

A science of existential risk, then, must involve delicate communication with the public and policy-makers.

4.5 Existential Risk as a Crisis Discipline

I have discussed a notion of creativity suitable for examination via the economic approach. One way of contextualizing such discussions is via a description of an epistemic situation. An investigation's epistemic situation is the sum of the challenges facing knowledge-generation, and the resources available for overcoming those challenges. I've thus far discussed the challenges facing paradigm X-risk investigations. Paradigm X-risks are *unique,* involve *wild systems,* and involve 2nd-order ignorance. Further, they are in the *public eye*, having the potential to generate over-reactions, a loss of faith, and mismanagement. These challenges are not insurmountable: many of them are not unique to X-risk, and so we can take our cue from other research areas.

It is useful to consider X-risk as a *crisis discipline*. In 1985, Michael Soule developed the notion by comparing conservation biology and cancer research. Both disciplines are geared towards a particular outcome (curing cancer, preserving biodiversity) and so membership in the crisis discipline turns on possession of a set of scientific expertise related to achieving that outcome. And so in addition to ecologists, conservation biology includes veterinary specialists, experts in land management, and so on. We've already had a hint about the wide variety of disciplines involved in X-risk—indeed, given our 2nd order ignorance it's actually unclear which disciplines will matter. In addition to being multi-disciplinary, crisis-disciplines are *normative*: X-risk is not simply in the business of describing or explaining low-probability, high impact events, but also in

ascertaining how to minimize the occurrence and impact of such events. A final similarity concerns the need to be tolerant of uncertainty:

> A conservation biologist may have to make decisions or recommendations about design and management before he or she is completely comfortable with the theoretical and empirical bases of the analysis (Soule 1985, 730).

In addition to having the characteristics of a crisis discipline, uniqueness and 2nd-order ignorance mean that X-risk studies will often occur in evidentially impoverished circumstances. I've analyzed similar epistemic situations occurring in 'historical sciences' such as paleontology, geology and archaeology (Currie 2018, 2016) and here, the success of the sciences is best explained by appeal to the speculative, creative nature of their approach (Alison Wylie has made similar arguments concerning archaeology, see Wylie 1999, Chapman & Wylie 2016).

I've sketched a set of investigative strategies which maximize evidential reach in historical science. Story-telling and scenario-building serve to maximize the empirical links between hypotheses. Historical reconstruction doesn't simply rely on the relationship between contemporary remains—traces—and the past, but on the connections between our hypotheses about the past. Further, such speculation often generates testable hypotheses (Currie 2017, Currie & Sterelny 2017): Historical scientists are highly creative in my sense. I've characterized historical scientists as 'methodological omnivores' (Currie 2015, 2018). Methodological omnivores engage in two distinctive behaviors. First, they construct epistemic tools and models calibrated to local context (as opposed to using general-purpose tools), enabling rich data to be generated. Second, a pluralistic, opportunistic attitude to techniques, methods and research approaches allows a wide range of perspectives, and different types of evidence, to be available. Finally, uniqueness can be mitigated by the use of partial analogies (Currie 2018, chapter 8).

Paradigm X-risks are not precisely in the same epistemic situation as conservation biology or paleontology—the tragedy of the uncommons is one difference—but nonetheless the

similarities can give us an inkling of the epistemic strategies which a science of X-risk should adopt. The science should be multi-disciplinary, pluralistic, opportunistic. Such a science very much meets the criteria for creativity in the sense I discussed in section 2. Each of these factors involve a community that does not pool, but rather explores solution space widely.

A science of X-risk, then, should be creative.

5. Discussion

A successful science of X-risk will be creative. But, as we've seen, contemporary scientific incentives don't often encourage creativity. Rather, they encourage cold-searches. Hence, the properties required for studying X-risk are not promoted in scientific communities. With this in place, I want to (1) characterize this problem abstractly: that science is 'badly-adapted' for studying X-risk; (2) continue my initial sketch of a well-adapted science of X-risk.

5.1 Well-Adapted Science

I've given reason to think that the incentive structures governing science are in a sense 'maladapted' for some epistemic situations, existential risk in particular. Where that situation calls for creativity, conservatism is encouraged. In this section, I'll characterise the problem abstractly and discuss its relationship with the economic approach on the one hand, and with work on the relationship between social values and science on the other.

The notion of 'well adapted' I want to develop concerns whether scientific incentives encourage the kind of work that is appropriate given an epistemic situation:

> **A scientific community is *well adapted* to the extent that the incentives of that community promote the attainment of desired research outcomes, given the epistemic situation at hand.**

Let's contrast being well-adapted in my sense—which is a relationship between a set of incentives and an epistemic situation—and the notion of an individual scientist or community being adapted to a set of incentives. Work in the economic approach is often not sensitive to this difference, and often the focus is more on the latter. Weisberg & Muldoon's landscape models and their descendants explore how different proportions of exploration strategies might be differentially optimal in various landscapes; Zollman's bandit models explore how scientists might learn to adapt to an epistemic situation. Here, an adapted scientist (or community) is one which maximises their returns (in terms of credit or knowledge) given some set of incentives. To be well-adapted in my sense, by contrast, requires those incentive structures to be themselves set in order to maximize the epistemic (or other) outputs that we desire given an epistemic situation. For instance, given the nature of X-risk, incentives in the community should encourage creativity. So, one sense of 'adaptive' concerns how well scientific behaviours maximise payoffs given a set of incentive structures. Another—mine—concerns how incentive structures might be organized to maximize epistemic outputs. Again, such a distinction is likely implicit in much of the economic approach, but it is useful to make it explicit.

Socially-inclined philosophers of science have argued that decisions about the pursuitworthiness of a scientific investigation or enterprise—what makes that research program a good one to do—turns on more than epistemic significance. Rather, a cost-benefit calculus is required to balance preferences for research outcomes, the efficiency of investigative approaches to those outcomes, as well as budgetary and ethical constraints. Philip Kitcher's approach is perhaps the clearest example (2001, 2011). For him, a science is 'well-ordered' to the extent that which research we pursue is decided by deliberation which approaches the cost-benefit calculus mentioned above (see also Cartwright 2006). So, discussion of the role of values in science often draws our attention to how the organization of science itself affects the efficiency of a research program: concerns about scientific organization and prioritization are taken as questions of resource distribution. Given a range of possible questions scientists might

be asking, by what principles should they direct their efforts? In short, a well-ordered science is one which balances (1) some suitably trained ('tutored') preferences, against (2) the efficiency of particular research programs in meeting those preferences, and (3) the costs (considered both in terms of finances, resources and ethics) of those programs[19].

Science might be well-ordered—that is, it might target the right programs in an efficient manner, but still not be well-adapted. Again, science is well adapted not when the scientists themselves adapt to the incentive structures in place (they'll do that well enough without our help) but when the incentive structure itself is conducive to the achievement of the goals we are interested in. The crucial contrast with Kitcher comes in the second and third aspects of his account of well-ordered science. The efficiency and cost of a scientific endeavour is not fixed, but determined in part by the social context in which the endeavour is carried out. And, to some extent, we have control of that social context.

The notion of a well-adapted science, then, brings two discussions into contact. First, considerations of how scientific communities react to epistemic situations. Second, considerations of the role non-epistemic values play in determining significance in science. A well-adapted science is one where the incentive structures are geared towards achieving the values discussed in the latter literature, and can do so in part in virtue of lessons from the former. On my view, understanding when a research program is well-adapted involves local, detailed work: thick descriptions of epistemic situations. The models favoured by the economic approach can play a critical role in suggesting and exploring potential interventions and effects.

5.2 A Science of Existential Risk

[19] In recent work, Kitcher further develops the notion of a well-ordered science, expanding the role of values in determining the significance of research questions, the role of the public in certifying scientific claims, and the importance of disagreement both within and without science. These developments don't affect the contrast with well-adapted science.

I've argued that, insofar as science doesn't promote creativity—hot-searching—it is not well set up for investigating existential risk. In the parlance of the last section, science is *badly adapted* to existential risk. However, researching existential risk is desirable: on the reasonable assumption that human extinction is a bad thing, just a little bit of knowledge which might lower the chances of extinction is going to be worth having. The question, then, is: how do we better adapt science to this epistemic situation? The crucial first step, I think, is to identify the sources of the maladaptation, and the second is to ask which of these we might do something about. I take myself to have gone some way towards the first part of this task. The second part, that is, identifying which aspects of the epistemic situation might be intervened on to better promote research outcomes pertaining to existential risks is tricky and, in this paper at least, above my paygrade. It is worth noting, however, that in the last sub-section I provided an explicit story for how such intervention strategies might be generated. Simplified models, the bread and butter of the economics approach, can create and explore hypotheses pertaining to the causes of conservatism and their possible interventions. For instance, O'Connor's (this issue) model should give us pause in assuming that increasing competitiveness will select for conservativeness. Combined with thick descriptions of epistemic situations, and perhaps integrated with empirical data (Harnagel, this issue) such models can then motivate trials of said interventions (Avin, this issue).

In section four I listed a set of factors which make X-risk demand a certain creativity in its scientists. And in section three, I listed a set of factors which make science non-creative. We should ask which of these features discussed in section three may be manipulated in such a way as to make a better adapted science.

My account of creativity involved a partial trade-off between exploration and efficiency. A science of existential risk should be exploratory, but science is geared towards efficiency and, as we've seen, at least some features promote efficiency at the expense of creativity (although I

doubt this is a necessary trade-off)[20]. Although paradigm X-risks face a particular epistemic situation, these are not unique insofar as there is bountiful overlap with other sciences. Above, I pointed out similarities between X-risk and sciences like paleontology and conservation biology. These sciences might provide inspiration for how to promote study of X-risk. And indeed some interventions have generally-speaking begun to be discussed and partly implemented. The National Science Foundation's 'transformation' grants explicitly attempt to fund exploratory research. Some scientific journals have adopted alternative publishing standards: PLOS ONE's policy of publishing any result which is judged to be methodologically sound is an example. And there are at least a few instances of alternative funding allocation strategies being trialled (Avin, this issue). I take my job in this paper to be making explicit the underlying reasons for wanting to explore these alternatives—particularly in light of X-risk—but exploring the space of solutions must be left for further work. Those challenges are summarized in table 1.

Table 1: Sources of scientific conservatism.

Property	Increases Conservatism by…
Peer Review (in funding/publishing)	• Tying success to pleasing peers. • Slowing down funding/publishing process.
Centralized Funding	• Tendency towards large projects. • Tendency towards safe projects.
Monistic publishing standards	• Only some results are published. • Bias towards research likely to produce those kinds of results.
'Public Eye'	• Possibility of miscommunication (either public overreaction or loss of faith).
Crowded Marketplace	• Scientists hunt out the safest bets in picking research directions.
Explicit success criteria in funding	• Makes exploratory research difficult to sell.

[20] Philosophers interested in the composition of scientific communities often argue that efficiency is achieved by some mixture of 'maverick' and 'follower' strategies (Kitcher 1990, Weisberg & Muldoon 2009, Thoma 2015), suggesting that there is not a tradeoff between creativity and efficiency. This may be right, but it is worth noting that my conception of creativity is not quite equivalent (as, properly speaking, it is a population-level phenomenon) and, as we've seen, such work is not geared towards understanding science's being well-adapted.

Disciplinary focus	• Interdisciplinary work/publishing detrimental to career (particularly early on).
Informal gate-keeping (gossip etc…)	• Reputation traps
Institutionalized teaching	• Tacit consensus
Differences in heritability of success between hot and cold labs.	• Lab-level selection for conservative strategies

I see this as a first-pass at a research agenda targeting the mitigation of conservation-causing features of science. This list is undoubtedly speculative, surely incomplete, and some features might be mischaracterized or misunderstood. But determining this will require further study, some of which might involve the economic approach, as well as the examination of case studies, and the kind of thick descriptions I have used here. And from this, it is plausible that further interventions might be trialled.

6. Conclusion

The essential tension in this paper is between aspects of science which make it *productive*—efficient—and those which make it *creative*. I doubt there is a clean trade-off between these virtues, but often we do need to decide whether scientific communities ought to be organized to favour productivity or creativity—cold or hot searches—and to what extent. And those decisions, I've suggested, should be made depending upon the epistemic situations those communities face. For X-risk, contemporary science is far too skewed towards productivity. A well-adapted science of X-risk, then, would be tailored towards generating creativity. I've provided a speculative, preliminary list of the sources of conservatism, and these deserve further study both via empirical and theoretical routes. Especially for cases such as X-risk, understanding how to create well-adapted science is urgent. However, whatever interventions we consider will likely be themselves speculative and risky: and these are risks being taken with the livelihood of individual scientists. In light of this, making such trials fair—providing safety nets, for instance—should be considered part of this research program as well.

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.

Avin, S. Wintle, B. Weitzdörfer, J. O hEigeartaigh, S. Rees, M. Sutherland, W. (forthcoming). Classifying Existential Risks. Futures.

Baron, M. G., Norman, D. B., & Barrett, P. M. (2017). A new hypothesis of dinosaur relationships and early dinosaur evolution. Nature, 543(7646), 501-506.

Baron, M. G., Norman, D. B., & Barrett, P. M. (2017). Baron et al. reply. Nature, 551(7678), E4-E5.

Baum, S. D., & Barrett, A. M. (2015). The most extreme risks: Global catastrophes. The Gower Handbook of Extreme Risk. Farnham, UK: Gower.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. (2017). Redefine statistical significance. Nature Human Behaviour, 1.

Boden, M. A. (2004). The creative mind: Myths and mechanisms. Psychology Press.

Bostrom N, Cirkovic MM (eds) 2008, Global Catastrophic Risks, Oxford University Press, Oxford

Bostrom, N. (2013). Existential risk prevention as global priority. Global Policy, 4(1), 15-31.

Bostrom N (2002). Existential risks: analyzing human extinction scenarios and related hazards. Journal of Evolution and Technology, vol. 9, no. 1

Bromham, L., Dinnage, R., & Hua, X. (2016). Interdisciplinary research has consistently lower funding success. Nature, 534(7609), 684.

Cartwright, N. (2006). Well-ordered science: Evidence for use. Philosophy of Science, 73(5), 981-990.

Currie, A & Avin, S. (forthcoming). Method Pluralism, Method Mismatch & Method Bias. Philosopher's Imprint.

Currie, A & Walsh, K. (2018). Newton on Islandworld: Ontic-Driven Explanations of Scientific Method. Perspectives on Science.

Currie, A., & Sterelny, K. (2017). In defence of story-telling. Studies in History and Philosophy of Science Part A.

Currie, A. (this issue). Introduction: Creativity & The Social Epistemology of Science. Studies in the History and Philosophy of Science.

Currie, A. (forthcoming). Geoengineering tensions. Futures.

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

Currie, A. (2016). Hot-Blooded Gluttons: Dependency, Coherence, and Method in the Historical Sciences. The British Journal for the Philosophy of Science, axw005.

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

Dasgupta, Patha, David, Paul A. (1994). Toward a new economics of science. Research Policy, 23(5):487–521.

De Cruz, H. (forthcoming). Prestige bias: An obstacle to a just academic philosophy. Ergo.

Douglas, H. (2009). Science, policy, and the value-free ideal. University of Pittsburgh Press.

Editorial (2017). Nature 550, 429.

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. American Sociological Review, 80(5), 875-908.

Gaut, B. (2010). The philosophy of creativity. Philosophy Compass, 5(12), 1034-1046.

Gaut, B. (2003). Creativity and imagination. The creation of art, 148-173.

Godfrey-Smith, P. (2008). Recurrent transient underdetermination and the glass half full. Philosophical Studies, 137(1), 141-148.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., ... & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. Proceedings of the National Academy of Sciences, 114(30), 7892-7899.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. Psychological review, 111(1), 3.

Isobe, H. (2016). Extreme Solar Flares as a Catastrophic Risk. Presented at Cambridge Conference on Existential Risk 2016.

Karieva, P. (forthcoming) Existential Risk due to Ecosystem Collapse: Nature Strikes Back. Futures.

Kitcher, P. (2011). Science in a democratic society. Prometheus Books.

Kitcher, Philip (2001), Science, Truth and Democracy. Oxford: Oxford University Press.

Kitcher, Philip. (1993) The Advancement of Science. New York: Oxford University Press.

Kitcher, Philip. (1990).The Division of Cognitive Labor. Journal of Philosophy, 87: 5–22.

Langer, M. C., Ezcurra, M. D., Rauhut, O. W., Benton, M. J., Knoll, F., McPhee, B. W., … & Brusatte, S. L. (2017). Untangling the dinosaur family tree. Nature, 551(7678), E1-E3.

Leonelli, S. (2016). Data-centric biology: a philosophical study. University of Chicago Press.

Liu, H. Y., Lauta, K. C., & Maas, M. M. (2018). Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research. Futures.

Longino, H. E. (2002). The fate of knowledge. Princeton University Press.

Mäki, U., Walsh, A., & Pinto, M. F. (Eds.). (2017). Scientific Imperialism: Exploring the Boundaries of Interdisciplinarity. Routledge. Forber, P. (2008). Forever beyond our grasp?. Biology & Philosophy. 135-141

Merton, Robert K. (1968). The Matthew Effect in Science. Science. 159 (3810): 56–63

Miller, J. D., & Felton, D. (2017). The Fermi paradox, Bayes' rule, and existential risk management. Futures, 86, 44-57.

Muldoon, R. (2017). Diversity, Rationality, and the Division of Cognitive Labor. IN: Boyer-Kassem, T., Mayo-Wilson, C., & Weisberg, M. (Eds.). Scientific Collaboration and Collective Knowledge: New Essays. Oxford University Press.

Muldoon, R. (2013). Diversity and the division of cognitive labor. Philosophy Compass, 8(2), 117-125.

O'Connor, Cailin. (this issue). The Natural Selection of Conservative Science.

O'Connor, Cailin. Brunner, Justin. (online first). Dynamics and Diversity in Epistemic Communities. Erknnetnis.

Paul, E. S., & Kaufman, S. B. (Eds.). (2014). The philosophy of creativity: new essays. Oxford University Press.

Pöyhönen, S. (2016). Value of cognitive diversity in science. Synthese, 1-22.

Price, H. (2016). The cold fusion horizon. Aeon (https://aeon.co/essays/why-do-scientists-dismiss-the-possibility-of-cold-fusion accessed 20/10/2017)

Shapin, S. (2008). The scientific life: A moral history of a late modern vocation. Chicago: University of Chicago Press.

Soulé, M. E. (1985). What is conservation biology?. BioScience, 35(11), 727-734.

Solomon, M. (2001). Social empiricism. Cambridge, MA: MIT press.

Stanford, P. K. (2015). Unconceived alternatives and conservatism in science: The impact of professionalization, peer-review, and big science. Synthese, 1-18.

Stanford, P. K. (2010). Exceeding our grasp: Science, history, and the problem of unconceived alternatives. Oxford University Press.

Strevens, Michael. (2013). Herding and the quest for credit. Journal of Economic Methodology, 20(1):19–34.

Strevens, Michael. (2003). The Role of the Priority Rule in Science. Journal of Philosophy, 100(2): 55–79.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. science, 331(6022), 1279-1285.

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

Torres, P (2017). Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks. Pitchstone Publishing (US&CA).

Tucker, A. (1998). Unique events: The underdetermination of explanation. Erkenntnis, 48(1), 61-83.

Wagner, Caroline S., Edwin Horlings, Travis A. Whetsell, Pauline Mattsson, and Katarina Nordqvist. "Do Nobel Laureates create prize-winning networks? An analysis of collaborative research in physiology or medicine." PloS one 10, no. 7 (2015): e0134164.

Wiener, J. B. (2016). The tragedy of the uncommons: On the politics of apocalypse. Global Policy, 7(S1), 67-80.

Weisberg, M. (2013). Modeling herding behavior and its risks. Journal of Economic Methodology, 20(1), 6-18.

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

Chapman, R., & Wylie, A. (2016). Evidential reasoning in archaeology. Bloomsbury Publishing.

Wylie A (1999) Rethinking unity as a "working hypothesis" for philosophy of science: how archaeologists exploit the disunities of science. Perspect Sci 7(3):293–317

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

Zollman, K. J. (2012). Social network structure and the achievement of consensus. Politics, Philosophy & Economics, 11(1), 26-44