A Mid-Level Approach to Modeling Scientific Communities

Forthcoming in Studies in History and Philosophy of Science

Audrey Harnagel

Abstract

This paper provides an account of *mid-level models*, which calibrate highly theoretical agent-based models of scientific communities by incorporating empirical information from real-world systems. As a result, these models more closely correspond with real-world communities, and are better suited for informing policy decisions than extant how-possibly models. I provide an exemplar of a mid-level model of science funding allocation that incorporates bibliometric data from scientific publications and data generated from empirical studies of peer review into an epistemic landscape model. The results of my model show that on a dynamic epistemic landscape, allocating funding by modified and pure lottery strategies performs comparably to a perfect selection funding allocation strategy. These results support the idea that introducing randomness into a funding allocation process may be a tractable policy worth exploring further through pilot studies. My exemplar shows that agent-based models need not be restricted to the abstract and the a-priori; they can also be informed by real empirical data.

1. Introduction

How should we organize scientific communities to produce significant, creative or revolutionary science? This foundational question is the object of inquiry for a growing cohort of philosophers of science (Weisberg and Muldoon 2009, Zollman 2007, 2010; Grim 2009, 2013; Alexander 2013, 2015; Holman and Bruner 2015, O'Connor and Bruner 2015). They use agent-based models to investigate questions like, 'how should scientists communicate to efficiently generate scientific knowledge?' (Zollman 2007, 2010) or 'what strategies should scientists use when selecting research approaches?' (Weisberg and Muldoon 2009). In this paper, I articulate a view of what these models could be capable of, and I apply my view by developing an agent-based model of science funding allocation.

Models are used for many purposes, and agent-based models of scientific communities are no exception. Diverse goals, whether they be explanatory, predictive, exploratory, or modeling for modeling's sake, motivate building and interpreting models differently. Within the philosophical literature, models are often interpreted as providing 'how-possibly' explanations. In this issue Currie argues that such models should be supplemented with 'thick' descriptions of scientific practice. I explore a different goal for computational models of scientific communities: providing predictive information about real-world communities, thus acting as a tool for informing policy decisions. In other words, my aim is to formulate a methodology to move from how-possibly towards more predictive models.

To develop this framework, I draw on Anna Alexandrova's methodology for building 'mid-level theories,' and adapt it to the modeling context (2017). Building a mid-level model involves

constructing the model in two directions: top down (starting with theoretical principles) and bottom up (starting with empirical literature). To date, modeling scientific communities within the philosophical literature has been dominated by theoretical approaches. Our increasing capacity to generate and analyze data about real scientific communities creates opportunities to integrate 'bottom-up' empirical approaches into these models. I construct a mid-level model, which utilizes bibliometric data, to investigate which funding allocation strategies maximize the generation of significant science.

I argue that using empirical research to set model parameters calibrates the model to regions of parameter space that have been observed in the real world, and as a result increases our confidence that the modeling results may obtain in the relevant target systems. However, the mid-level model remains abstracted and idealized in certain respects, and so the model alone does not guarantee its predictions or the effects of actual interventions; rather it serves as a tool for discriminating between different policy interventions to be explored further through pilot tests (see Figure 2, below).

In this paper, then, I have two aims. First, detailing a process for generating *mid-level models* by incorporating empirical information into high-level theoretical models. Second, providing an exemplar of a mid-level model by extending high-level models of science funding allocation (Avin 2017). The results of my model show that allocating funding by modified lottery systems generates comparable results to other funding methods. These results support the idea that introducing randomness into funding allocation processes may be a constructive policy worth exploring further through pilot studies. Modified lottery allocation systems reduce the influence of conservative biases on what gets funded, which in turn could promote more creative science. My exemplar shows that agent-based models need not be restricted to the abstract and the a-priori, they can also be informed by real empirical data. As a result, these models can be employed to investigate policy interventions.

I begin with a brief introduction to the existing philosophical work using agent-based models of scientific communities, which motivates moving towards more predictive models. Next I develop my concept of mid-level models using the example case of modeling science funding systems. In Section 4 I present my model results. I discuss the interpretation of my model results and argue that they support a call for pilot studies in Section 5.

2. Agent-Based Models of Scientific Communities

One influential approach to modeling scientific communities is to represent scientists as agents who explore an 'epistemic landscape,' an abstract representation of scientific problem space. Weisberg and Muldoon (2009) developed the concept of an epistemic landscape based on evolutionary biologists' concepts of fitness landscapes. In their model the landscape consists of a three dimensional grid; each 'patch' on the grid corresponding to an x,y coordinate represents a research 'approach,' which is a catchall term for how a scientist investigates a topic. A given approach could include questions, methods, background theory etc. For example, different approaches to studying a child's ability to engage in pretend play (a single research topic

represented by an epistemic landscape) include investigating how children play with peers versus adults or studying the differences between individual and group play (Weisberg and Muldoon 2009, 228). Distance between coordinates is a measure of similarity between points, where closer points are more similar. The third dimension, height, of Weisberg and Muldoon's landscape corresponds to the 'epistemic significance' of the results generated by pursuing a particular research approach. Some scientific discoveries are more influential than others, and more influential discoveries are represented by greater height. The variability in epistemic significance of scientific results determines the topography of the landscape. The original landscape used by Weisberg and Muldoon (2009) consists of a flat landscape with two peaks, one higher than the other (Figure 1). Many modelers of scientific communities have drawn on the concept of an epistemic landscape, and different models incorporate new features or more complicated landscapes (Grim 2009, Grim et al. 2013, Alexander et al. 2015, Thoma 2015, Muldoon 2013, Avin 2017, Pöyhönen 2016). In section 3 I discuss my model, which uses a similarly structured epistemic landscape.

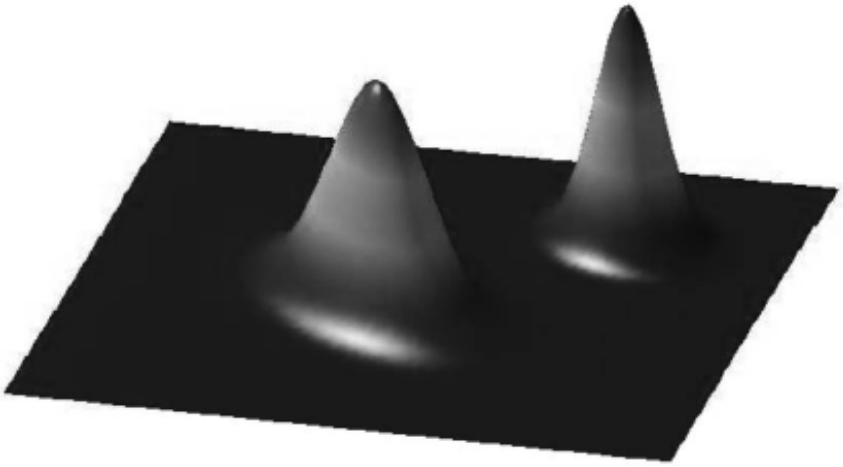


Figure 1: Weisberg and Muldoon 2009.

One criticism of this approach to modeling scientific communities is that it is unclear how these highly abstracted and idealized models map onto real-world target systems (Avin 2014, Rosenstock et al. 2016). Why think the actual epistemic landscape of science is adequately represented by two Gaussian peaks? And even though it most likely is not, what kind of information can be gleaned from a two-peak model to inform the organization of complex, real-world scientific communities? There is a rich philosophical literature on how models, although never perfectly accurate descriptions of real phenomena, can provide information about their target systems (Teller 2001; Suppes 19670a, 1960b; Suppe 1977, 1989; van Frassen 1980; Giere 1979; Carwright 1983). By the end of the 20th century, similarity accounts of model-world relationships, which allowed flexibility for the modeler to determine how similar a model must be to a target system for her purposes, had gained traction (Teller 2001, Giere 1979, Cartwright 1983). There is now general consensus within philosophical modeling literature that *which similarities* between model and target system are relevant, and *to what degree* they are similar

depends on the case at hand and the goals of the modeler (Parker 2009, Teller 2001, Weisberg 2013, Grim et al. 2013, Suarez 2003, Bhakthavatsalam and Cartwright 2017).

If that is the case, perhaps the high degree of abstraction and idealization built into agent-based models of scientific communities is not a problem for modelers' purposes. Agent-based models are often taken to provide 'how-possibly' stories where the model shows that some phenomenon could possibly result from the (often simplistic) conditions in the model (Rosenstock et al. 2016). In a how-possibly model there is no need to show that the modeling results are likely, or that they do in fact obtain in the world; the results are merely *possible*. Other authors have advocated the use of agent-based models as formalized thought experiments (Avin 2014, Currie and Avin, 2017). While these are productive uses for such models, my project takes a different approach. I am interested in what it would take to build models that can provide predictions about real-world communities, and as a result can be used as tools for policymakers to investigate potential policy interventions.

The conclusions that can be drawn from existing models fall short of this goal. The main shortcoming of agent-based models is eloquently summed up by Rosenstock et al. (2016): "We do not have a good sense of which real world communities are well represented by which epistemic …models" (17). As a result, we cannot be sure whether real-world epistemic communities correspond with the areas of parameter space that are represented in the model "or some other area, or some other models with different assumptions" (17). To use models to inform (at least in part) policy interventions, we need models that support stronger claims than the results are merely *possible*. We want reason to believe that the model results will obtain in the world with some non-trivial probability. While these high-level models give us valuable conceptual tools, like the epistemic landscape, they underdetermine the specific features that must be included to apply modeling results to a policy context. One way to support stronger claims is by shrinking the parameter space represented by these models, so we are more confident that the parameters of the model do in fact represent conditions that are encountered in real-world target systems. This can be accomplished by incorporating empirical information about scientific communities to calibrate model parameters.

Similar to the gap between highly abstracted models of scientific communities, and real-world communities themselves, Alexandrova observes a disparity between philosophical theories of well-being, and practical measurements and conceptions of well-being used to define policy. Alexandrova's answer is to use real-world practice and empirical studies to shape more operational, context-dependent theories about well-being, which she calls 'mid-level theories.' In the next section, I use Alexandrova's approach as inspiration to develop a mid-level model of science funding allocation, which incorporates empirical information to help narrow the gap between abstracted models and real-world scientific communities.

3. A Mid-level Model of Science Funding

Philosophical theories, like models, often generalize and abstract from any specific context. As a result, theories alone are often not detailed enough to provide a practical guide for application; rather, a 'rulebook' is required to apply a theory in a policy context (Alexandrova 2015, 228-229). To address this problem, Alexandrova develops the concept of a mid-level theory.

A mid-level theory lies between a general and abstracted 'high' theory and specific measures in practical and scientific contexts (2017, xxviii). To construct such a theory requires working in two directions: from the top down (starting with high theories) and from the bottom up (starting with the existing empirical base) (xiii). I refer to my model of science funding allocation as a 'mid-level model' because, similarly, it falls between the kinds of high-level models described in Section 2 and a practical policy context, or a controlled pilot experiment to test a particular policy intervention (see Figure 2). To construct such a model I take a similar top down and bottom up approach.

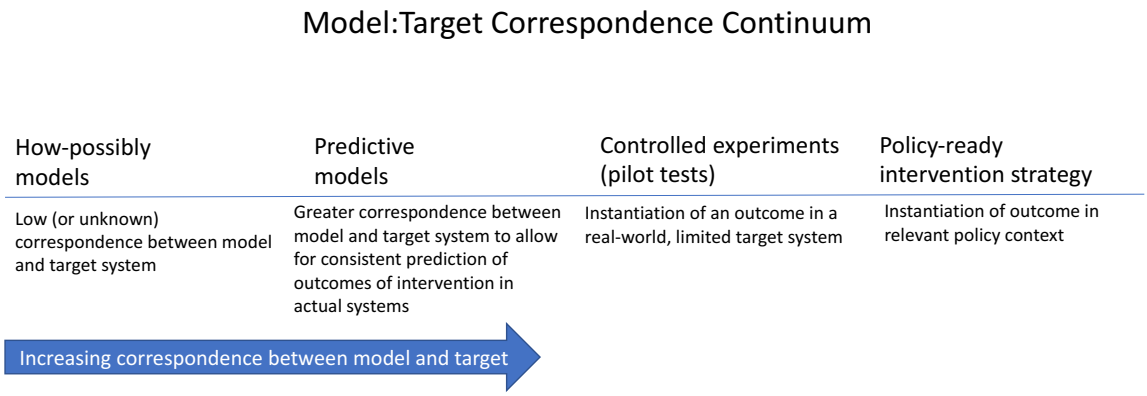


Figure 2

I start with a high-level agent-based model of science funding developed by Avin (2017). Avin's model provides the structure for two essential features of science funding systems: 1) a distribution of scientific projects 2) mechanisms for selecting which projects to fund. Avin draws on the concept of an epistemic landscape described in Section 2 to represent a distribution of grant applications with varying potentials for generating significant science. However, Avin's model cannot answer questions necessary to develop specific policy proposals. One particularly salient problem is - how can we know whether the theoretical epistemic landscapes have any bearing on what the epistemic landscape of science actually looks like? To address this question, I take a bottom up approach, using the distribution of citation counts from scientific publications as a proxy for scientific significance. I draw on additional empirical literature assessing the performance of peer review, the major method for allocating research funding, to inform the mechanisms for grant selection in my model. To describe my model in greater detail,

I first outline Avin's high-level model, followed by the bottom-up modifications I made to adapt it to a mid-level model.

3.1 Building from the top down: a high-level model of science funding allocation

Avin (2017) builds on previous epistemic landscape models, and adds the additional element of funding allocation by a centralized funding body. The basic structure consists of an epistemic landscape similar to Weisberg and Muldoon's, where a two-dimensional grid represents a research topic with a range of possible projects, and a third dimension (height) represents the significance of the result obtained upon successful completion of the project. The 'ruggedness' of the landscape is determined by randomly generating a specified number of Gaussian hills, which are scaled according to a specified maximum height. Avin's landscape is significantly more rugged than Weisberg and Muldoon's two-hill landscape, however the topography of the landscape remains randomly generated, which makes it difficult to tell whether this landscape is in fact a good representation of scientific problem space.

The agents in Avin's model represent scientists investigating a research topic. The particular x,y coordinate of an agent on the landscape represents the project that agent is currently pursuing. When agents are generated, each one receives a countdown counter, which indicates the time remaining until the project is completed. Upon completion, an agent contributes the significance of its own project to a sum of the collective significance generated by all of the agents. The agent then moves to the highest neighboring patch on the landscape, where its 'neighborhood' is defined as the 3x3 square centered on its current position.

An additional variable of Avin's model is the 'visibility' that agents contribute to as they explore the landscape. This visibility determines how well funding bodies can estimate the significance of a given project. As agents are generated on the landscape and move around, they contribute vision of their local area to the total vision of the landscape by the centralized funding body.

The significance of projects changes over time in Avin's landscape, which represents a departure from previous landscape models. When a scientist completes a given project (represented by a specific patch in x,y space), the significance of that patch is reduced to zero indicating the 'winner takes all' priority-rule governing scientific discoveries (Strevens 2003). Additionally, the novelty of similar projects in the local area around the location of the discovery is reduced upon completion of a project. Finally, when an agent completes a project, a new hill is generated at a random location on the landscape, simulating how scientific discoveries open new directions for research.

Funding allocation is simulated by selecting which agents within an application pool will receive funding to continue research. The agents that do not receive funding are removed from the landscape. A population of new agents is also generated each round to join the pool of existing candidates for funding. Selection proceeds according to five different funding mechanisms:

Best:
Selects candidates located at the highest points on the landscape, regardless of the visibility of their locations. This mechanism selects the most promising projects from god's eye perspective, and therefore serves as an upper bound for the productivity of the landscape.
Best visible:
Filters out the candidates located at 'invisible' locations (located too far away from present or past projects for their location to be visible). The candidates on the highest locations from the remaining visible areas are selected. Avin claims that this strategy gets closer to representing selection by peer review, but remains idealized because it assumes that a selection panel has gathered all available information from all the different agents past and present when making selection decisions.

Lotto:
Randomly selects candidates from the selection pool regardless of visibility or height of their locations.

Oldboys:
According to this mechanism there is no selection. The candidates that were originally initiated on the landscape continue to explore for the duration of the simulation.

Triage:
This strategy provides some funding based on knowledge of the visible parts of the landscape, and distributes the remainder of the funding randomly outside of visibility. In the model, 50% of funding is allocated by the 'best visible' strategy within visible regions, and 50% is allocated by 'lotto' outside of visible regions.

Avin's results indicate that increasing the dynamics of the landscape by changing the significance of a given research approach over time results in a similar performance of 'best,' 'lotto,' and 'triage' strategies. My model supports a similar result, where increasing the dynamics of the landscape decreases the difference in performance between the various strategies. Avin observes that the 'vision range' of scientists has an impact on the epistemic success of peer-review-based mechanisms. Greater 'vision' of the landscape results in higher success of the best visible strategy relative to the others. In Avin's simulations, however, the 'best visible' strategy is often outperformed by 'triage' and 'lotto' strategies.

These results foster interesting food for thought, especially for promoting creativity in scientific exploration. The models suggest that the conservative bias of peer review, which Avin models as a limitation in vision of the epistemic landscape, can hinder the ability of the scientific community to successfully explore the landscape and generate significant scientific knowledge. However, Avin's model remains limited for the purposes of predicting whether one funding mechanism will perform better than another in a real system. While Avin's model improved on extant landscape models in important ways, namely altering landscape significance over time, many of the features of Avin's model remain arbitrarily assigned. Additionally, Avin's approximation of the performance of peer review is generated by the simple mechanism of

limiting vision of the landscape. It is unclear whether this process well represents the performance of peer review. Following the mid-level framework, I sought out empirical literature that would help me adapt and concretize the features of Avin's model in a more data-driven way.

3.2 Building from the bottom up: incorporating empirical information

My model draws on two sources of empirical data to calibrate two distinct features of the model. First, I use scientific publication data to construct an epistemic landscape. Second, the funding mechanisms in my model draw on data from social scientific studies of the performance of peer review.

Rather than randomly generating a topographical landscape, I focused on constructing a landscape on the basis of current empirical measures of scientific significance. The obvious choice was citation data from scientific publications. While my landscape remains idealized in certain respects, it represents distributions of scientific significance (measured by number of citations) across scientific disciplines and over time that are found in the world. Citation metrics are not a perfect measure of the significance of science, but they provide an approximation of the topography of the landscape of scientific knowledge that is more realistic than arbitrarily located Gaussian hills on a Euclidean grid. Citation counts generated by a given scientific discipline vary from year to year, so my landscape is also dynamic over time. Using this data-driven approach, I adopt the concept of an epistemic landscape form a high-level model, and improve its correspondence to the real world scientific problem space.

Constructing an epistemic landscape based on publication data requires an algorithm for mapping publications spatially, which clusters publications according to similarity in references and keywords. Happily, computer scientists working for the University of San Diego (UCSD) have constructed such a map (Börner et al. 2012). The map is populated by a sample of publication information from the Thompson Reuters Web of Science database ranging from 2006-2016. Using the UCSD Map of Science, I developed a landscape where scientific subdisciplines (such as Human Evolution or Aerospace) are oriented in 2D space based on the conceptual similarity between subdisciplines (Figure 3).[1] Each subdiscipline is represented by a node, and rather than constructing a continuous grid, my landscape is made up of this collection of discontinuous nodes. The size (diameter) of each node depends on the number of papers published in that subdiscipline in a given year. The more papers published in a given field, the larger the node. The subdisciplines can be aggregated into 13 larger disciplines (Figure 4).

The third dimension of the landscape is determined by citation count. Citation counts remain the only standardized, quantitative and ubiquitous metric for the significance of a scientific

[1] By using references and keywords as the raw material for clustering subdisciplines, this strategy takes into account both similarity in conceptual content between two disciplines and required skills. Keywords will pick out similar techniques even if they are used in conceptually different disciplines. Also references citing papers with similar methods even if they are across conceptually different subdisciplines are factored into the similarity index.

publication, and therefore represent the only metric that would allow for systematic empirical input into the model. While developing better measuring sticks for scientific value is an avenue for future work, for the purposes of this model, citation counts provide an accessible first approximation of scientific significance (Sinatra et al. 2016, Lehmann et al. 2006, Owens 2013). The height of a node corresponds to the average citation count (significance) of the scientific publications that populate that subdiscipline on the landscape for a given year. Figures 3 and 4 show a 2D representation of the landscape, where the height dimension is represented by color. The color of the node indicates this significance like a heat map, which ranges from blue (high significance) to red (low significance) (Figure 3).

Figure 3

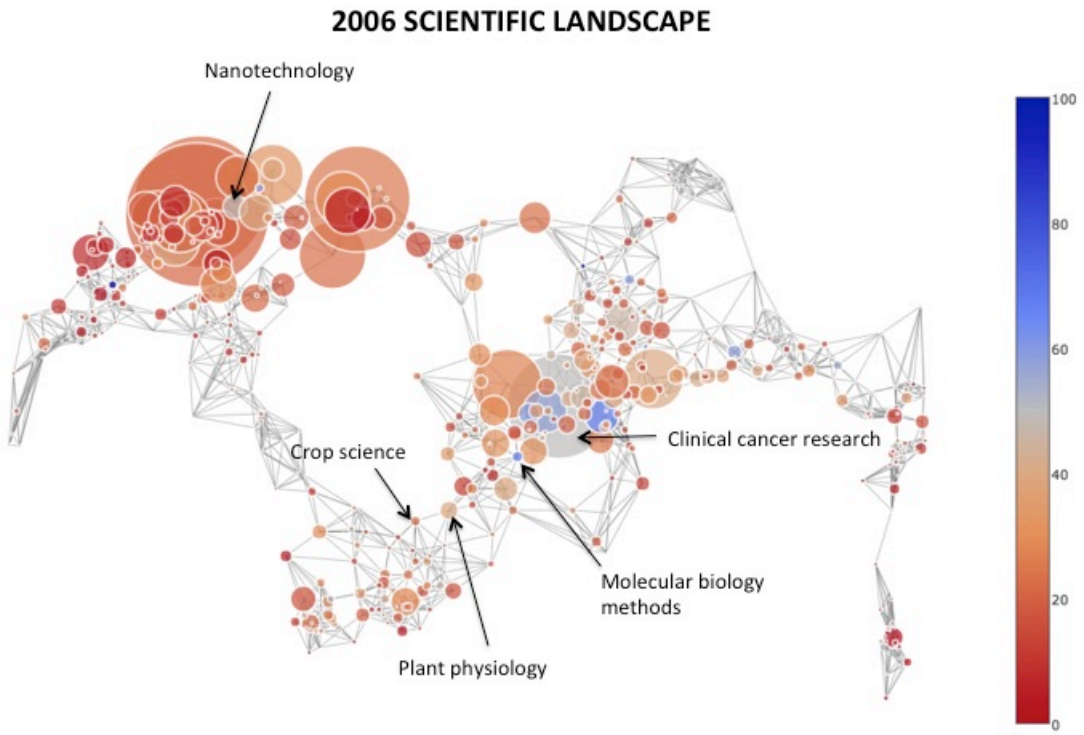
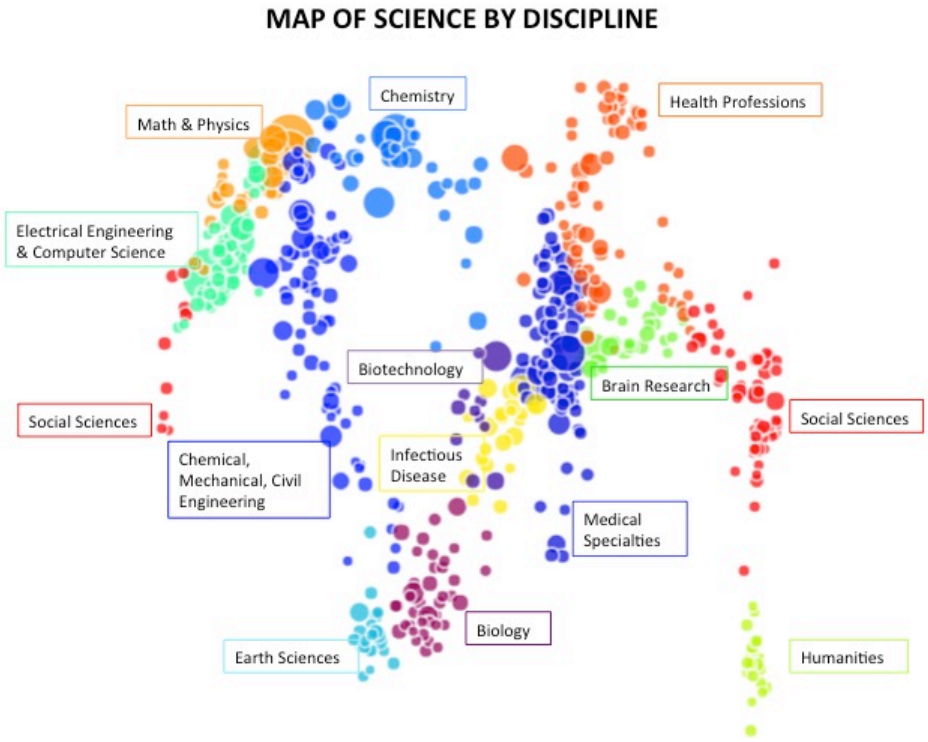


Figure 4



The epistemic landscape is used to generate a distribution of potential scientific projects; various different mechanisms are then used to select which projects receive funding.[2] The mechanics of this process are similar to Avin's model. At the beginning of the model each agent (scientist) is a candidate for funding, and its location represents the scientist's field of research. Once initialized on a particular node (subfield), an agent applies for a grant. The quality of the grant proposal is determined by the 'height' of the subdiscipline the scientist belongs to. Not all scientists working within a subdiscipline generate results of the same significance (accumulate the same number of citations). Therefore to introduce heterogeneity within a given node, the significance of a particular grant proposal is drawn from a single-exponential Poisson distribution, where a parameter of the distribution is the mean citation count for that node. Statistical analysis shows that this probability law well represents the distribution of citation counts of scientific papers in a variety of disciplines (Vieria and Gomes 2010). A selection mechanism is then applied to the pool of proposals. Similar to evolutionary models, the agents whose proposals receive funding 'survive' to conduct their project, and the agents who do not receive funding disappear from the landscape. In each successive year of the model, a new round of funding is carried out. My simulation proceeds according to the same process as Avin's with a few modifications:

1) In my model, agents are randomly assigned a size-weighted node; therefore agents are more likely to be assigned to a larger node (subdiscipline with more publications) than a smaller node. This differs slightly from Avin's model where agents are randomly generated anywhere on the landscape.
2) Upon completing a project, Avin's agents seek out the neighboring patch with the greatest significance, and relocate to that patch for the next round. In contrast, in my simulations the agents do not move after they are initiated on the landscape. This is largely due to the difference in interpretation of the locations on the landscape between the two models. On Avin's landscape, and most other epistemic landscape models of scientific communities, a patch on the landscape represents a particular research approach (project). Therefore scientists move from project to conceptually related project over the course of a career, and ultimately explore the landscape in a step-by-step fashion. On my landscape, an individual node represents a particular subdiscipline of science. While some scientists transition between subdisciplines over the course of their career, it is certainly more rare and costly to jump between different nodes on my landscape. If the landscape encompassed a smaller scope, for example a particular

[2] My model assumes that these two features are independent – the topology of the landscape is independent of funding bodies' decisions of what to fund. This is problematic for funding bodies that play a key role in determining the future directions of scientific research by allocating large percentage of the resources invested in scientific research. Therefore, my model should be interpreted as representing a marginal funding body whose resources contribute a small percentage of the global investment in science (not enough to drastically change the landscape of science itself).[3.1] Examples of this type of funding body abound: public funds contributed by a small national government (e.g. Denmark), small private foundations or specific institutes within larger national funding agencies such as the National Institute of Allergy and Infectious Disease (NIAID), an institute within the NIH. Notably, it would be challenging to use a data-driven approach to generate such models since we only have data from the funding decisions that were in fact made. Modeling alternative funding decisions, and the subsequent evolution in response to them, would remain a theoretical pursuit.

[3.1] I am grateful to Shahar Avin for suggestions on interpreting my model given this limitation.

discipline or subdiscipline of science, it might be appropriate to model scientists' exploration of the landscape in comparable ways to extant work on epistemic landscape models. For the scope of my model, once an agent is initiated on a node it remains on that node for the duration of the time it has grant funding.

After each round of the simulation, the landscape is updated with publication data from the consecutive year, which means that the landscape evolves over time. Certain subdisciplines grow or shrink over time; the number of citations generated in a given subdiscipline also evolves over time, so the height of the nodes is dynamic.

I model four different strategies for funding allocation, which are described below. In developing these funding strategies, I adopted ideas from Avin's funding mechanisms (my strategy 'Perfect Selection' is the same as Avin's 'best' strategy, and our 'Lotto' strategies are identical). At the outset of this project I intended to compare simulations of peer review to alternative funding strategies. I turned to social science literature on the performance of peer review to calibrate the parameters of my modeled funding mechanisms to findings from empirical research. I was surprised by the dearth of research on the performance of peer review, particularly with regards to how well grant peer review scores predict future bibliometric outcomes.[3] Within the small body of literature on this topic, there is lack of consensus on the performance of peer review, but a growing body of evidence suggests that peer review has a limited ability to discriminate between proposals that will ultimately generate higher impact publications and those that will not (Danthi et al. 2014, 2015; Doyle et al. 2015; Fang et al. 2016; Kaltman et al. 2014; Van den Besselaar and Sandstrom 2015, Lauer et al. 2015, Reinhart 2009, Pier et al. 2017). Lauer found that grants with higher percentile rankings according to grant review scores generated highly cited articles with a correlation "slightly better than chance" (95% confidence interval 0.51-0.53) (240). Reinhart concluded that assessments by the research council "have little predictive power for the future publication success" (806). Based on the existing literature, one could interpret Lotto to be a proxy for the performance of peer review. That is not to say that peer reviewers cannot discriminate between good and bad applications; rather current paylines at most funding institutions allow only a small percentage of applications to receive funding, therefore peer reviewers are being asked to discriminate the very cream of the crop from a pool with abundant meritorious applications.

The state of the literature indicates that we are in the early stages of understanding the performance of peer review as an indicator of future bibliographic success. As a result, I do not attempt to explicitly simulate the performance of peer review. Instead I compare a Perfect Selection strategy (an obvious overestimate of the performance of peer review) to two modified lottery funding methods, Triage and Focal Randomization, as well as a purely random Lotto strategy. Even without explicitly simulating peer review, the mere marginal difference between the performance of Perfect Selection and modified lottery strategies in my simulations

[3] This gap has also been noted in the recent literature review by Guthrie et al. (2017).

indicate that peer review may not perform better than lottery-based strategies on a dynamic landscape.

Perfect Selection:
This strategy is the same as Avin's 'best' strategy – it selects the agents (grant applications) with the highest significance values on the landscape. This provides an upper bound for the epistemic potential of any given pool of grant applicants.

Focal Randomization:
Empirical research suggests that there is a high degree of variability in ranking of a pool of grant applications within funding panels. One study found that across 45 different funding panels, only 9% of grant applications were always funded and 61% were never funded (Graves et al. 2011). This left 29% of applications that were sometimes funded and sometimes not depending on the make up of the particular research panel. One potential method for allocating funding proposed by Elise Brezis automatically funds or rejects, respectively, projects that are unanimously agreed to be top or bottom ranking proposals by all reviewers (Brezis 2007). The applications for which there is disagreement among reviewers are funded by lottery.

The Focal Randomization strategy in my model combines Graves et al.'s results with Brezis's proposed model. Grant applications are ranked from most significant to least, and within this rank-ordered list, the top 9% of applications are held constant at the top of the ranking and 61% of grants at the bottom of the ranking are held constant. The middle portion of the list is then randomly shuffled and applications are funded from the top until funding runs out. Focal Randomization is thus a modified lottery strategy that combines quality assessment with random allocation.

Triage:
This strategy is modeled after various funding agencies' processes of rejecting a certain portion of applications after an initial quality review. Depending on the funding body, sometimes this cursory review involves a simple check to see if an application meets the funding body's eligibility criteria, or it may involve a more sophisticated filtering process. In my model, applications are rank ordered, the bottom 50% of candidates are rejected, and the remaining 50% are randomly allocated funding until the available funding runs out. This is loosely based on the process used by the NIH where, following expert review and rebuttal, the bottom 50% of applications are removed from consideration. My 'triage' mechanism is similar to Avin's in the sense that it represents a modified lottery strategy that combines an assessment of the potential of a given project based on its position on the landscape with lottery allocation.

Lotto:
Proposals in the applicant pool are selected at random (same as Avin).

These funding mechanisms foster creativity in different degrees. Currie (this issue) develops a definition of creativity that contrasts 'cold search' and 'hot search' approaches for exploring an epistemic landscape. In a cold search, agents methodically search their local solution space,

relying on priors for what kinds of approaches might be fruitful. Peer review, Currie argues, fosters a cold-search approach where projects that gain the approval of peers are often consistent with previous work and priors of the scientific community. In contrast, a hot search puts less weight on priors, which allows hot-searching agents to 'jump' about a landscape and try things that may not be consistent with what has been done in the past. In my model, funding strategies based solely on the significance of the node that the agent occupies (Perfect Selection) are more likely to promote cold searches, because funding is based on which areas of science are generating the most significant results at the time in which funding is allocated. Additionally, this type of approach will concentrate funding in prosperous nodes, rather than spreading funding across the landscape. Strategies that incorporate a random component (Focal Randomization, Triage, Lotto) are more likely to promote 'hot search' approaches, since funding allocation is not solely dependent on which areas of science have been previously generated significant results. A random (or partially random) allocation method is also more likely to distribute resources across the epistemic landscape. In this sense, increasing the randomness of funding allocation is one way to increase the creativity of science on the population level.

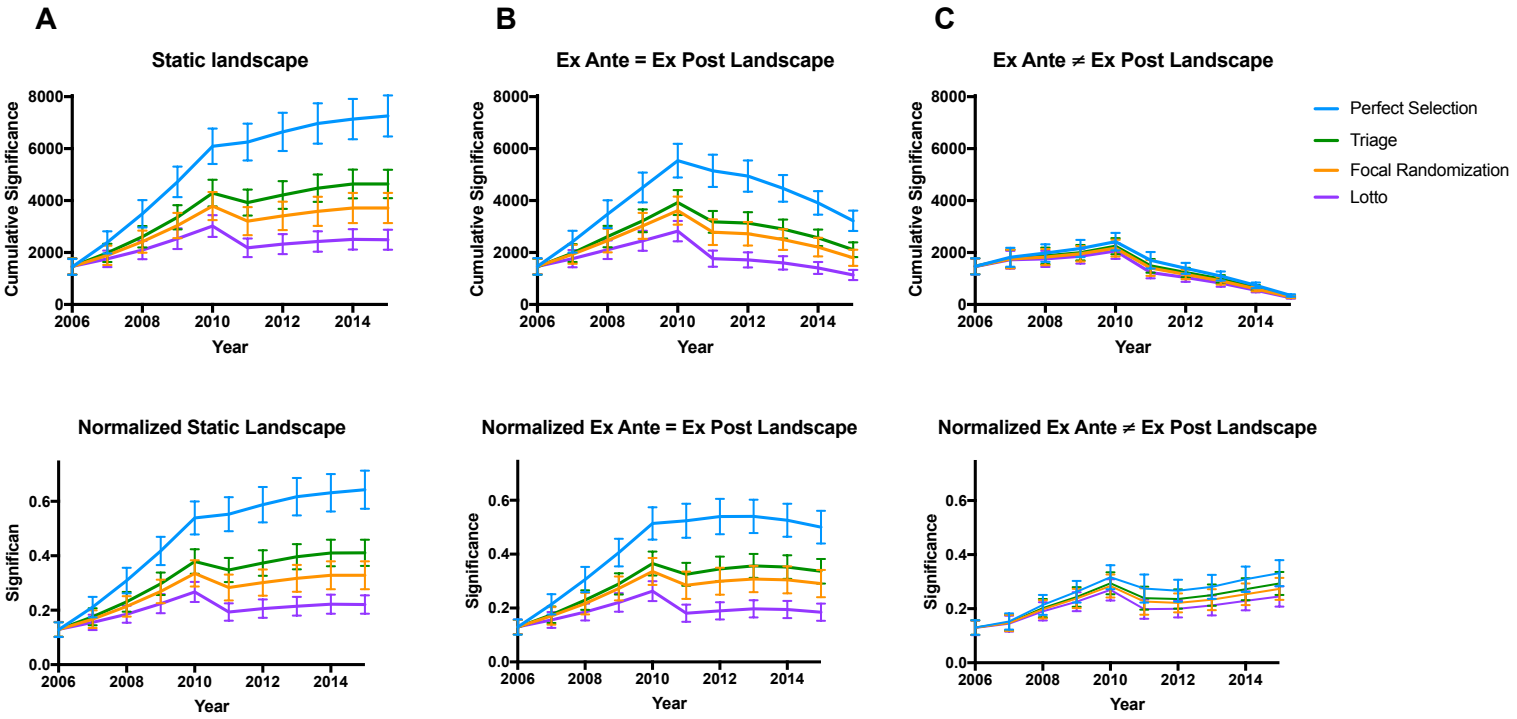
4. Model Results

In the following analysis, I present plots of the yearly significance generated by the scientific community over 10 years of simulation. I vary the dynamism of the scientific landscape, and present results from three different conditions: 1) a static landscape (no change in significance over time); 2) a landscape where the significance of a project remains constant from funding allocation to publication; 3) a landscape where the significance of a project could change between funding allocation and publication. When I incorporate these landscape dynamics over time, the difference in cumulative significance generated by the various strategies collapses. On a dynamic landscape, the difference in performance between random allocation by Lotto and Perfect Selection is only marginal.

In Figure 5, the top panel shows the total significance generated by 300 agents simulated over 10 years. The data point for each year represents the sum of the significance of every project completed in a given year i.e. the projects that ran out of grant funding and subsequently published results. The error bars represent the standard deviation across 2000 repetitions of the simulation.

The bottom panel shows the total significance generated normalized by the total citation count that constitutes the landscape for each year. Because the epistemic landscape is constructed from relatively recent publication data, the total number of citations making up the landscape decreases over time, since more recent publications have not had time to generate as many citations as older publications. Citation counts typically stabilize approximately 10 years post-publication.

Figure 5



The panels from left to right show the three different dynamic conditions. The results in panel A were generated on a static landscape. The landscape generated from publication data in 2006 was simulated for a 10-year duration. Agents applying for a grant in 2006 encounter the same landscape as agents applying for a grant in 2010. In this condition, neither the significance nor the size of subdisciplines changes over the course of the simulation.

Panel B shows a simulation where the significance of a given project remains static from the time it is funded and assigned to the time it is published (countdown = 0). In my model this is a range of one to five years depending on the random assignment of the duration (countdown) of the grant. Upon finishing a project and publishing results, a scientist 'cashes in' on the significance value that was assigned when the project was funded some years before. This condition assumes that the value assigned to a scientific project *ex ante* when it is funded remains unchanged when the *ex post* results are published.[4] This condition is more dynamic than a purely static landscape because when an agent finishes a project and applies for a new

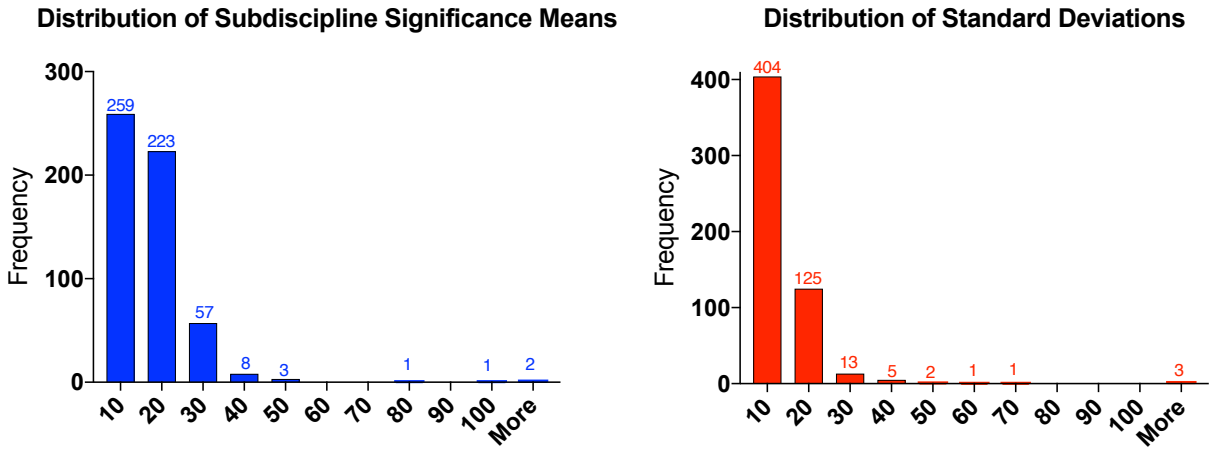
[4] Because in the Ex Ante = Ex Post condition (Condition B) the significance of a project is assigned when the project is funded, not at the time of publication, these results cannot simply be normalized by the total landscape significance on the year of publication. Instead, the year in which funding was assigned for each project that is published, and the corresponding cumulative significance of the landscape for that year. The sum of these landscape significance values is then divided by the number of published projects, and this average provides the normalization constant for Condition B.

grant, the landscape will have evolved since the agent was initiated on the landscape. Under this condition, agents applying for grants in 2006 encounter a different scientific landscape than agents applying for grants in 2010. Another way to conceptualize this is that for each agent, the landscape remains static for the duration of a project, but changes before the inception of a new project.

Panel C shows a simulation where a project receives an *ex ante* assignment of value that is used as an input to the selection mechanism to determine whether or not the project is funded. This *ex ante* value is assigned based on the significance of the agent's node at the time they apply for the grant. The projects that receive funding and go on to completion are then assigned a new *ex post* significance value at the time of publication (at the end of the duration of the grant). This *ex post* value is determined by the significance of the node at the time of publication, some years after the initial value was assigned. The likelihood that the *ex post* value of a project is similar to the *ex ante* value depends on how variable the significance of a given subfield (as determined by citation metrics) is over time. On a landscape where significance does not vary much over time, the *ex post* value will likely be similar to the *ex ante* value; thus the simulation results would look similar to the two previous conditions.

However, the simulation results and analysis of the scientific landscapes used in these models show considerable variation in significance of individual subfields over a 10-year time span. Figure 6 shows the distribution of the mean significance of the 554 subfields that make up the landscape over the span of 10 years, and the distribution of the standard deviation in significance across the 10-year period.

Figure 6



The long tail in the distribution of standard deviations indicates a high degree in variation in the topography of the landscape over the ten-year period. Additionally the distribution of the standard deviations of the subfields is similar to the distribution of the means, indicating a high

degree of variability in the dataset. Notably, the level of dynamism is modeled using data-driven changes in the significance of subdisciplines measured by citation counts; it is not an arbitrary rate of change. Figure 5 reveals that when the dynamics of the scientific landscape are in play, differentiation between the various funding strategies collapses.

The order of performance for the five strategies remains the same across the conditions and regardless of whether the cumulative significance is normalized. Perfect Selection marks the upper bound.[5] The next best performing methods, Triage and Focal Randomization, represent modified lottery strategies. Lotto, the lowest performing strategy, had the highest degree of randomness. The most striking result is that introducing the dynamics of the scientific landscape in condition C collapses the relative performance gap between Perfect Selection and the remaining strategies such that Perfect Selection performs only marginally better than Lotto. This indicates that even if grant reviewers could evaluate proposed projects with 100% accuracy based on the epistemic landscape at the time when funding was allocated, upon publication, the projects selected based on this perfect evaluation are only marginally more significant (generate more citations) than a pool of randomly selected projects. This result is clear from comparing conditions A and B to condition C in the normalized plot. Examining the normalized data shows that the collapse in the performance of the various funding strategies is not due to variations in the number of citations that populate the landscape from year to year. This collapse is due to the fact that the projects with the highest significance when funding was allocated were no longer the most significant when results were published. In the model context, this is due to the changes in significance of a given subdiscipline over time.

It may not be surprising that if significance changes unpredictably over time, a random method will perform just as well or better than any other method for choosing what to fund. It is surprising, however, that the scientific landscape changes to such a degree on the time scale of the duration of an individual grant that other strategies catch up to Perfect Selection. In other words, the model reveals that landscape is more dynamic than we may have thought, which impacts the likelihood of success of strategies that rely on prediction or randomness respectively.

In Avin's simulation, the strategies 'best', 'lotto' and 'triage' perform similarly when the landscape is dynamic over the course of the simulation. The consistency between Avin's results, which are based on robustness analyses of a purely theoretical model, and my results, based on a model that tracks empirical data, further bolster Avin's conclusions. My model shows that

[5] I note that in the *ex ante* significance $\neq$ *ex post* significance (condition C) simulations, the Perfect Selection marks the upper bound under the assumption that funding allocators cannot predict the future. In other words, Perfect Selection represents the results that would be generated if funding bodies allocated resources based on a perfect estimation of the significance of the project *at the time* that funding was allocated. A different allocation strategy, call it the 'soothsayer' strategy, where funding bodies could predict which projects would result in maximal significance *at the time of publication*, could generate greater significance over the duration of the model.[3.1] While an interesting strategy to pursue in future models, this represents an even more optimistic view of the ability of funders to assess the potential of scientific projects *ex ante* than Perfect Selection.

[3.1] I am grateful to Adrian Currie for suggesting a 'soothsayer' strategy.

some of the effects observed in Avin's high-level model hold up on a landscape that is more realistic to real world scientific problem space.

My simulations indicate that selecting projects based on the state of the world at the time of funding may not be an optimal strategy for generating significant results in the future due to the dynamic nature of the epistemic landscape. The charge of peer review is not so much to assess a scientific proposal for its feasibility and impact in the current context, but to predict the significance of the proposed research in a future context. My model reveals how dynamic the epistemic landscape of science is seeing as the significance of a given subfield can change substantially in a short amount of time. Given limited information about the future epistemic landscape of science, it is unclear whether peer reviewers and funding bodies can accurately anticipate which projects will be significant in the future. Evidence from empirical studies about the predictive performance of peer review suggests that peer reviewers have difficulty predicting *ex ante* which scientific projects will have high *ex post* impact. As a result, it may be fruitful to experiment with funding allocation systems that intentionally introduce randomness into the funding allocation process, such as modified or pure lottery strategies.

5. Interpreting the Model

There are many things my model is not. It is not explanatory. It is also not necessarily predictive of how various funding strategies will perform relative to one another in a real system. It does not, in and of itself, justify policy intervention. My model does, however, give us better reason than high-level theoretical models to believe that various types of lottery strategies might perform equal or better than current mechanisms of peer review. Further improvements to my model, consistent with a mid-level modeling approach, could allow for this model to be useful for justifying pilot experiments in particular policy contexts. I will spend the remainder of this section defending these claims.

The question I set out to address in my mid-level model is: which funding allocation strategies maximize the generation of significant science? As a first pass, all that is required of a model is to measure the success of various strategies relative to each other, such that an institutional designer or policymaker could pick an optimal strategy to investigate further. Notably, this type of model aims for prediction rather than explanation. Answering my question of interest does not require an explanation of *why* a given strategy performs better (although this is perhaps a worthy goal to pursue eventually); it need only predict which system performs the best. As a result, the features of my model need not map onto causal structures, at least not in a way that the model-users (including myself) specifically understand.[6] Any additional insights into the internal mechanisms of the system gained through modeling are an added bonus. For the purposes of discriminating between one policy over another, we really want models that predict outcomes. Thus far, extant high-level models do not satisfy this predictive goal.

[6] It is hard to imagine a predictive model that does not include representations of *any* causal factors. I am merely saying that we need not be able to demonstrate which features are causal and by what mechanisms.

One particularly thorny challenge for understanding the effectiveness of scientific funding strategies based on *ex ante* assessments is missing counterfactuals. It is difficult to measure how well science funding systems are working because we do not have counterfactual knowledge of the outcomes that would have resulted if the proposals that were rejected had instead received funding. Proposed projects that do not receive funding are almost never carried out, so any measure of success is already limited to only those projects that receive funding. We do not know what the epistemic landscape looks like for unexplored projects, and so it does not influence the landscape in my model.

This blind spot is not unique to my model. It confronts all empirical work that strives to measure scientific outcomes. Additionally, this is not a data collection problem – these counterfactuals are unknowable. There is no way to measure the success of a scientific project that is never carried out. Short of funding all proposed projects, the problem of missing counterfactuals will pervade any type of empirical study on the outcomes of funding structures. The only way to explore such counterfactual situations is to construct theoretical epistemic landscapes. Given the limitations of theoretical landscapes discussed on Section 2, I argue that there is valuable information to be gleaned from modeling a subset of all possible alternatives of funding allocation (the subset of projects that were in fact funded).

Whether or not the results of the model hold true in the real world must be confirmed by testing them. I do not claim that my model alone provides confirmation that the phenomena observed in the model will happen in the real world.[7] I adopt Alexandrova's *open formulae* view of models, where a model serves as a tool for sorting through many potential hypotheses and articulating some to test in the real world. Mid-level models can be used to inform which pilot tests or controlled experiments to execute before making more sweeping policy decisions, thereby moving from less to more correspondence along the Model: Target Correspondence Continuum (Figure 2). Finally, mid-level models are certainly not the *only* information a policymaker should use to justify policy reform. Using families of models, some theoretical and some empirically based, may be the best strategy to address complex policy problems like science funding allocation. Other factors such as economic assessments, political considerations, etc. will continue to play a role in driving policy. In many policy contexts, models offer valuable tools to investigate the many variables of complex systems in a systematic, structured way; they can therefore provide one piece of justificatory evidence towards intervention.

The question remains, why build a model in the first place instead of skipping straight to pilot tests? Pilot experiments, while smaller scale than a major policy overhaul, are costly in a number of different ways. Obviously there are costs in terms of time and resources required to set up various funding structures, run them and analyze results, but there are additional political and opportunity costs (see Avin, this issue). In the science funding case, piloting new

[7] Models themselves (particularly idealized models) cannot provide this confirmation because it is often unknown whether given assumptions in models are satisfied in their target systems, or in some cases there are certain assumptions that are *known to be unsatisfied* in the target system (Alexandrova 2008).

funding allocation strategies sends a message that the current strategy may be sub-optimal, which challenges the credibility of the institutions allocating funding. There is also an opportunity cost when testing different strategies for funding allocation. If strategy A is better at allocating funds to significant projects than strategy B, significant projects in application pool B may be passed over due to a flaw in the funding allocation strategy. While these costs on a small scale (such as a pilot experiment) may be justified with the larger goal of improving the whole system of funding allocation, stakeholders still need some reason to believe that the strategies to be tested in pilot studies have the potential to yield better results than current systems. Mid-level models can provide some reason to believe that a given policy intervention will improve the status quo, and, in concert with other knowledge, generate hypotheses to test in pilot studies.

Surely there is a spectrum of better or worse hypotheses that models could be used to generate. We want reason to believe the hypothesis resulting from the model will obtain in the world (with some non-trivial probability), thereby justifying the proposal that we should test that hypothesis. In other words, when informing policy decisions we want models to yield *plausible* outcomes, rather than merely *possible* ones. By using empirical data about scientific communities to calibrate the parameters of my model, I shrink the parameter space represented by the model to areas that are actually observed in the world under some circumstances. This takes a step towards addressing Rosenstock et al.'s concern that we do not have a good sense of which areas of parameter space represented in the model correspond with real-world epistemic communities (17). Using empirical information as a guide allows us to be more confident that the model parameter space does in fact correspond to conditions encountered in real-world communities. This helps move from how-possibly models with a vast parameter space towards more predictive models, which have greater potential to predict outcomes in real-world target systems (Figure 2).

Grim et al. argues that to be confident in simulation results, we need assurance that the aspects of a model "claimed to correspond to reality really do" and the "aspects purported not to correspond are as innocuous as they are posited to be" (2013, 2378). The authors claim that lack of correspondence between a model and target constitutes failure only if the model leaves out relevant variables, and "only if reality is more complicated in ways relevant to the goals of the simulation" (2383). One way to avoid the question of whether non-corresponding features are innocuous is to make those features corresponding. Oftentimes this comes at the cost of simplicity; however, there are cases when increasing the correspondence of a model does not markedly increase its complexity, particularly with regards to setting parameters. For example, in my model, the height of a specific location on the landscape is still represented by a single value, but rather than choosing a random value for this height (as in Avin's model), the range of significance values is empirically grounded. In this way, incorporating empirical results increases the correspondence between relevant features in the model and features of the real-world system, and also minimizes non-corresponding features. We have some reason to believe that features built into the model do in fact obtain in the world because they have been observed empirically. Because there are many assumptions built into the model that are inconsistent with real systems, however, models must be validated through real-world tests.

The epistemic landscape is one feature of models of scientific communities that can be vastly improved with the introduction of empirical information. The landscape in my model represents a measurement of what the collective scientific community constitutes significant science. This contrasts with arbitrarily generating landscapes with abstract characteristics of the distribution of scientific significance (such as areas of greater and less significance, significant projects tend to cluster together, etc.). However, there are ways in which the correspondence between my model and real-world communities could be further improved. For instance, the sparse empirical literature assessing the performance of peer review over time limits the extent to which a peer review allocation system can be accurately modeled (Guthrie 2017). A better understanding of how well peer review selects projects that lead to significant outcomes is necessary to effectively estimate how alternative strategies may perform relative to the status quo.

The application of a model is context dependent, and different applications will require models built for different scales. My model takes all the subdisciplines of science as its scope, but a finer grained bibliometric analysis could allow for more precise landscapes at different scales (such as specific disciplines e.g. biology, or more niche subfields e.g. invertebrate neuroscience). Depending on who was applying this model, many parameters could be further calibrated by empirical knowledge about the target system of interest. How many applicants to include in the pool, and what subdisciplines to include could be set according to which funding system the modeler was interested in. Since science is highly globalized, and the scientific publication community is largely international, I would argue that citation metrics remain a reasonable measure of scientific significance regardless of scale. This metric could also be further refined based on context, however, to account for more local measurements of what constitutes significant science. For instance, if developing new technology is an important goal of the funding body being modeled, then it may be appropriate to rate the significance of methods papers greater than research articles.

Fostering collaborative relationships between philosophers, scientists and policymakers would facilitate the construction of mid-level models. Scientists and policymakers offer specialized local knowledge and resources useful for the bottom-up approach to modeling. As Alexandrova advocates, a more integrated division of labor between philosophers and people closer to the empirical work (scientists, policymakers, etc.) would yield theories that are conceptually well-founded but also sensitive to practical constraints of measurement and use (2017, xxxi). For a model to be helpful in an applied policy context it must also be sensitive to practical constraints.

Returning to my model, whether the results provide sufficient confidence to justify initiating pilot tests of lottery-based funding allocation strategies will depend on the particular context in which the model is applied, and the assessment of the policymakers using the model (in concert with other sources of information). Introducing randomness into the funding allocation process separates funding allocation from the preconceptions of peer reviewers, and thereby would likely reduce conservative bias. This is especially important when reviewers are tasked with

identifying the excellent proposals from a pool of high quality applications, which is the case peer reviewers face when funding rates are less than 10%, as is often the case for many funding agencies today (Pier 2018). Faced with this situation, bias and random variation between reviewers' assessments of merit will play a more influential role in ultimate funding decisions. This contrasts to a situation where funding rates are higher (say 50%) in which peer reviewers merely have to discriminate the top half of the pool from the bottom half. In the latter case, it is more likely that less subjective epistemic rationales between reviewers will enable the separation of applications that receive funding and applications that do not (Pier 2018, Graves 2011).

For the purposes of fostering novel and creative science, introducing some degree of randomness into funding allocation presents an attractive way forward. Avin (2017) and Currie (this issue) each provide reasons for thinking that peer review increases the conservatism of science, which may be justified in some contexts if a more conservative strategy also increases the efficiency of science – that is, if it allowed scientists to uncover significant results faster. However, my results and Avin's show that modified lottery strategies do not result in large losses of efficiency when the epistemic landscape is dynamic. This suggests that science could afford to be more creative, and as Avin emphasizes, less costly, without sacrificing scientific productivity. Given that some funding bodies are already exploring these types of funding schemes (see Avin, this issue), my model provides further evidence that such initiatives are justified. Mid-level modeling results may provide confidence for additional funding bodies to consider similar policies, particularly funders interested in fostering novel, pioneering science.

6. Conclusion

Taking stock of the philosophy of science modeling literature, agent-based models provide promising tools for better understanding complex and dynamic scientific communities. However, existing models are limited in their applicability to real-world systems, which hinders the ability of such models to provide actionable policy recommendations. Questions of how to organize scientific communities to maximize the generation of significant science lie at the core of social epistemology. With the right tools, philosophers of science could be poised to weigh in on policy-relevant problems such as, 'Which funding strategies are optimal?' This paper develops a modeling approach to move towards this goal.

Building on existing theoretical models, I propose constructing 'mid-level' models that incorporate empirical data to calibrate parameters to real world phenomena, thereby increasing the correspondence between model and target system. One of the drawbacks of idealized models is that many assumptions are required to capture a complex system, and whether certain assumptions are satisfied is often unknown. One way of limiting this set of unknowns is by setting parameters to empirically observed values. This is especially attractive when empirical data can be incorporated without further complicating the model structure. Mid-level models seek to provide insights about target systems by representing them with a degree of accuracy that warrants real-world intervention. What degree of certainty this requires and how it is achieved will vary depending on the target system and the goals of model

users. Broadly, however, I argue that mid-level models usually strive for prediction over explanation. For instance, when designing funding schemes, my model aims to provide information about which funding allocation strategies will yield the best outcome relative to the other strategies – it does not seek to explain what causal structures are responsible for one strategy performing better than another. For many policy-relevant problems, comparing the performance of different interventions to each other is an adequate resolution of prediction to justify moving on to pilot tests.

My model of science funding strategies demonstrates that mid-level modeling is not merely an aspirational framework for future modeling endeavors, but a strategy that can be employed by coupling accessible data and extant or new high-level models. My model draws on Avin's epistemic landscape model to build a mid-level model from the top down. To build from the bottom up, I utilized bibliometric data from the common publication database Thompson Reuters Web of Science, and data from published social science literature on peer review. By detailing the process I used to construct my model, I hope to inspire others to consider what kinds of empirical data could be harnessed to inform model parameters. A data-driven approach is not appropriate for all questions philosophers of science wish to explore with agent-based models, but for those interested in informing policy, mid-level models provide more actionable results than current theoretical approaches.

Acknowledgements

I am grateful to Shahar Avin for his guidance in both developing the ideas and writing the code for the model presented in this paper, and for consulting with me across an ocean. I also owe many thanks to Adrian Currie, Stephen John, Breke Harnagel, Matt Kalmans and two anonymous reviewers for helpful comments on earlier drafts. Thanks also to Lauren Reeder for statistical assistance. A version of this work was presented at the 2017 European Philosophy of Science Association (EPSA) conference at the University of Exeter, and comments there helped shape this paper. This publication would not have been possible without the support of the Thouron Award.

References

Alexander, J. M. (2013). Preferential attachment and the search for successful theories. *Philosophy of Science, 80*(5), 769-782.

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. (2008). Making models count. *Philosophy of Science*, *75*(3), 383-404.

Alexandrova, A. (2015). Well-being and Philosophy of Science. *Philosophy Compass*, *10*(3), 219-231.

Alexandrova, A. (2017). *A philosophy for the science of well-being.* Oxford University Press.

Avin, S. (2014). Breaking the grant cycle: on the rational allocation of public resources to scientific research projects. Unpublished PhD Thesis. University of Cambridge Department of History and Philosophy of Science.

Avin, S. (2017). Centralized Funding and Epistemic Exploration. *The British Journal for the Philosophy of Science*.

Bhakthavatsalam, S., & Cartwright, N. (2017). What's so special about empirical adequacy?. *European Journal for Philosophy of Science*, 1-21.

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., ... & Boyack, K. W. (2012). Design and update of a classification system: The UCSD map of science. *PloS one*, *7*(7), e39464.

Brezis, Elise S. Focal randomisation: An optimal mechanism for the evaluation of R&D projects. *Science and Public Policy* 34.10 (2007): 691-698.

Cartwright, N. (1983). *How the laws of physics lie*, Clarendon Press, Oxford.

Currie A., and Avin, S. (2017). Method pluralism, method mismatch and method bias. Manuscript submitted for publication.

Currie A. (2018). Existential risk, creativity and well-adapted science. *Studies in History and Philosophy of Science Part A*, this issue.

Danthi, N.S., Wu, C.O., Shi, P., and Lauer, M.S. (2014). Percentile ranking and citation impact of a large cohort of National Heart, Lung and Blood Institute-funded cardiovascular R01 grants. *Circulation Research*, *114*(4), 600-06.

Danthi N.S., Wu, C.O., DiMichele, D.M., Hoots, W.K. and Lauer, M.S. (2015). Citation impact of NHLBI R01 grants funded through the American Recovery and Reinvestment Act as compared to R01 grants funded through a standard payline. *Circulation Research, 116*(5), 784-88.

Doyle, J.M., Quinn, K. Bodenstein, YU.A., Wu, C.O., Danthi, N.S., and Lauer, M.S. (2015). Association of percentile ranking with citation impact and productivity in a large cohort of de novo NIMH-funded R01 grants. *Molecular psychiatry, 20*(9), 1030-36.

Fang, F. C., Bowen, A., and Casadevall, A. (2016). NIH peer review percentile scores are poorly predictive of grant productivity. *Elife*, *5*, e13323.

Giere, R. (1979). *Understanding scientific reasoning*. Hold, Rinehart and Winston, New York, Second edition, 1984.

Graves, N., Barnett, A. G., & Clarke, P. (2011). Funding grant proposals for scientific research: retrospective analysis of scores by members of grant review panel. *BMJ*, *343*, d4797.

Grim, P. (2009). Threshold phenomena in epistemic networks. In *AAI Fall Symposium: Complex Adaptive Systems and the Threshold Effect,* 53-60.

Grim, P., Rosenberger, R., Rosenfeld, A., Anderson, B., & Eason, R. E. (2013). How simulations fail. *Synthese*, *190*(12), 2367-2390.

Guthrie, S. Ghiga, I. and Wooding, S. (2017). What do we know about grant peer review in the health sciences? *F100Research.*

Holman, E. and Bruner, J.P. (2015). The problem of intransigently biased agents. *Philosophy of Science, 82*(5), 956-968.

Kaltman, J.R., Evans, F.J., Danthi, N.S., Wu, C.O., DiMichele, D.M., and Lauer, M.S. (2014) Prior publication productivity, grant percentile ranking and topic normalized citation impact of NHLBI cardiovascular R01 grants. *Circulation Research, 115*(7), 617-24.

Lauer, M.S., Danthi, N.S., Kaltman, J., and Wu, C.O. (2015). Predicting productivity returns on investment: thirty years of peer review, grant funding and publication of highly cited papers at the National Heart, Lung, and Blood Institute. *Circulation Research, 117*(3), 239-243.

Lehmann, S., Jackson, A.D., and Latrup, B.E. (2006). Measures for measures. *Nature*, *444*(7122), 1003-1004.

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

O'Connor, C. and Bruner, J.P. (2015). Dynamics and diversity in epistemic communities. Manuscript.

Owens, B. (2013). Research assessments: Judgment day. *Nature*, *502*, 288-290.

Parker, W.S. (2009). II - Confirmation and adequacy-for-purpose in climate modeling. *Aristotelian Society Supplementary Volume*, *83(1)*, 233-249.

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., ... & Carnes, M. (2018). Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences*, *115*(12), 2952-2957.

Pöyhönen, S. (2016). Value of cognitive diversity in science. *Synthese*, *194*(11), 4519-4540.

Reinhart, M. (2009). Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. *Scientometrics, 81*(3), 789-809.

Rosenstock, S., O'Connor C., and Bruner, J.P. (2017). In epistemic networks is less really more? *Philosophy of Science, 82(2),* 234-252.

Sinatra, R., Wang, D., Deville, P., Song, C., and Barbasi, A.L., (2016) Quantifying the evolution of individual scientific impact. *Science*, *354(6312)*, aaf5239-1-aaf5239-8.

Strevens, M. (2003). The role of the priority rule in science. *Journal of Philosophy, 100,* 55-79*.*

Suarez, M. (2003). Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science, 7,* 225-244.

Suppe, F. (Ed.) (1977). *The structure of scientific theories*. Chicago: University of Illinois Press.

Suppe, F. (1989). *The semantic conception of theories and scientific realism.* Chicago: University of Illinois Press.

Suppes (1960a). A comparison of the meaning and use of models in mathematics and the empirical sciences. *Synthese, 12*, 287-300.

Suppes (1960b). Models of data. In E. Nagel and PO. Suppes (Eds.), *Logic, methodology and the philosophy of science: Proceedings of the 1960 international congress* (pp. 251-261). Stanford: Stanford University Press.

Teller, P. (2001). Twilight of the perfect model model. *Erkenntnis*, *55*(3), 393-415.

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

Van den Besselaar, P., and Sandstrom, U. (2015). Early career grants, performance, and careers: A study on predictive validity of grant decisions. *Journal of Informetrics, 9*(4), 826-38.

van Frassen, B.C. (1980). *The scientific image*. Oxford: Oxford University Press.

Vieira, E.S., and Gomes, J.A. (2010). Citations to scientific articles: Its distribution and dependence on the article features. *Journal of Informetrics, 4*(1), 1-13.

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

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world.* Oxford University Press.

Zollman, K. J. (2007). The communication structure of epistemic communities. *Philosophy of Science, 74*(5), 547-587.

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