Evaluating Formal Models of Science

Mike Thicke

September 2, 2018

1 Abstract

This paper presents an account of how to evaluate formal models of science: models and simulations in social epistemology designed to draw normative conclusions about the social structure of scientific research. I argue that such models should be evaluated according to their representational and predictive accuracy. Using these criteria and comparisons with familiar models from science, I argue that most formal models of science are incapable of supporting normative conclusions.

2 Introduction

"How do we best design social institutions for the advancement of learning? The philosophers have ignored the social structure of science. The point, however, is to change it." -Philip Kitcher, "The Division of Cognitive Labor" (1990)

This Marx-inspired rallying cry from Philip Kitcher is perhaps the founding statement of what Alvin Goldman calls "systems-oriented social epistemology" (Goldman 2011, p. 18). According to Goldman, systems-oriented social epistemology is concerned with the "social practices, procedures, institutions, and/or patterns of interpersonal influence that affect the epistemic outcomes" of "epistemic systems." As Kitcher suggests, the ultimate aim of this branch of social epistemology is to make normative recommendations about the organization of scientific research.

In recent years there has been a proliferation of formal models of science purporting to make such normative claims. These models include economics-inspired equilibrium models like Kitcher's that model scientists as maximizing fixed utility functions (Strevens 2003), Bayesian network models where scientists adjust beliefs by observing nearby peers (K. Zollman 2007), and agent-based simulations where scientists traverse abstract "epistemic landscapes" (Weisberg and Muldoon 2009). In general, I take a formal model of science to be an abstract mathematical structure that purports to represent some aspect of scientific inquiry. The behaviours of these models are taken to suggest corresponding behaviours in actual scientific practice that may be more or less desirable according to some standard. Frequently, though not always, these models take the form of computer simulations containing “agents” that behave according to simple sets of rules meant to be similar to ways in which scientists behave.

Despite the proliferation of such models, there has been little systematic discussion of how to evaluate them. While philosophers such as Hands (1997) and economists such as Mirowski (2004) have criticized the importation of economic methodologies into philosophy of science, with Kitcher’s models as prime exemplars, few have addressed the practice of developing *models* of scientific activity specifically. Given that papers based on formal models of science are frequently published in high prestige journals such as *Philosophy of Science* and *Synthese*, garner significant citations, and are invoked to support a wide range of arguments, such a discussion is long past due.

Rare instances of such general discussion of formal models of scientific inquiry include Martini and Pinto (2017) and Reijula and Kuorikoski (2018). Martini and Pinto argue that current models cannot be assumed to correctly predict the behaviour of scientific communities without connecting them to empirical data. Reijula and Kuorikoski argue that to be useful models of science need to fulfil three criteria. First, their conclusions should not be obvious. Second, they should be based on “clear theoretical ideas”. Third, their results should depend not on arbitrary assumptions but on assumptions with empirical support. Many current models of science, they argue, fail to fulfil one or more of these criteria.

I agree with both of these assessments of the current state of formal models of science. I hope that my paper can add to their analyses by examining more closely *why* these criteria are important and *why* empirical data is important for assessing these kinds of models. I do this by examining philosophical accounts of model evaluation in science, especially economics.

Ultimately, I will propose that formal models of science be evaluated according to their predictive and representational accuracy. After demonstrating how these criteria can be used to evaluate familiar models in science, such as global climate models, Schelling’s racial segregation model, and Arrow and Debreu’s general equilibrium model, I will apply them to evaluating formal models of science. A central premise of this approach is that models in philosophy ought to be evaluated by the same standards that philosophers apply to models in other fields. By that standard, I will argue that the vast majority of formal models of science, including the most well-known examples (K. Zollman 2007; Weisberg and Muldoon 2009; Hong and Page 2004), are incapable of supporting their explicit aim to make normative claims about the organization of scientific research. Consequently, they must either pursue different aims or adopt methodologies that establish more significant links between models and their targets.

This article is aimed at an audience already familiar with at least some instances of formal models of science. This article does not explain any particular model in depth, but instead discusses aspects of those models to illustrate and support arguments about the practice in general. For those unfamiliar with formal models of science, I recommend first reading some of the model papers referenced above, as well as the discussions of Martini and Pinto (2017) and Reijula and Kuorikoski (2018).

3 What are the aims of formal models of science?

The explicit aims of formal models of science are very much in line with Kitcher's mission statement. Weisberg and Muldoon's epistemic landscape model purports to suggest that "a mixed strategy where some scientists are very conservative and others quite risk taking leads to the maximum amount of epistemic progress in the scientific community" (Weisberg and Muldoon 2009, p. 227). K. Zollman (2007) concludes that, "In circumstances where speed is very important or where we think that our initial estimates are likely very close to the truth, connected groups of scientists will be more reliable. On the other hand, when we want accuracy above all else, we should prefer communities made up of more isolated individuals"(586). Hong and Page (2004) conclude that, in some conditions, "a random group of intelligent problem solvers will outperform a group of the best problem solvers," which "provides insights into the trade-off between diversity and ability"(16389).[1] According to Romero (2016), his model shows that, "self-correction [in science] is a fragile property: once we move away from the utopia and consider less utopian scenarios, the procedure of aggregating experimental evidence by meta-analysis can easily lead communities of frequentist scientists astray"(66). He then proceeds to suggest "institutional interventions" that could improve science's ability to self-correct. These examples are representative: formal modellers are explicitly concerned with offering assessments of and prescriptions about the institutional structure of science.

While assessment and prescription are clearly the central goal of such models, they are not exclusive. Strevens (2003), for instance, argues that his model of different reward structures shows that, because for scientific research normally only the first instance of discovery is socially valuable, the priority rule is often better at dividing cognitive labor than other potential reward schemes, such as rewarding scientists proportionally to their marginal contribution to solving a problem. This argument is similar in spirit to the above normative claims, but Strevens emphasizes the explanatory function of his model over its normative function. His model, he says, *explains why* science employs the priority rule despite its apparent excesses (Merton 1957). Similarly, Weisberg and Muldoon (2009) could be interpreted as an explanation for *why* diversity is important in science. The implication of this shift in emphasis is that we already accept that the priority rule and diversity are essential features of science, and we are interested in discovering why they are essential. Underlying Strevens' argument is the assumption that science is collectively rational: because the priority rule persists, there must be some good reason for it. Since his argument presupposes the rationality of science, there is no prospect for reforming it based on the results of his analysis, and therefore there could be a lower evidential threshold

[1] Unlike the other papers discussed here, Hong and Page (2004) do not discuss *science* specifically, but rather *problem solving* generally. However, due to the similarity of their approach to that employed by philosophers of science, I treat this paper as part of the formal models of science literature.

for accepting the implications of his model.

Martini and Pinto (2017) argue that, in addition to their normative aims, at least some models have a descriptive aim: “to represent, or describe, the target systems better than other (analytic or just non-quantitative) models” (231). For instance, they point to Weisberg and Muldoon’s (2009) claim that an advantage of their agent-based simulation is that it is better able to represent scientists with limited access to information than the models of Kitcher (1990) or Strevens (2003). Indeed, increased representational ability or “realism” is frequently cited as an advantage of one modelling approach over another. Thoma (2015b) argues, for example, that her model is superior to Muldoon and Weisberg’s because it allows scientists to be more or less flexible in their approaches to a problem and because her modelled scientists don’t engage in pointless and unrealistic duplication of effort (Thoma 2015a, p. 463).

Another possible aim of formal models of science is to serve as engines of inquiry rather than as directly normative or descriptive. Rosenstock, Bruner, and OConnor (2017) argue, for instance, that Zollman’s models are best seen as providing “how-potentially stories”: “They direct our attention to phenomena that might, potentially, occur in real communities” (235). That is, formal models could identify possible features of science that merit further investigation through other means. An example of this function can be seen in Perović et al. (2016). Discussing the loosely-connected experimental groups at Fermilab, they observe: “Such networks seem similar to the kind of loose connectedness that Zollman’s simulations suggest are optimal, rather than to the wheel structure or the maximally connected networks” (9). This claim is not epistemically dependent on the representational accuracy or normative claims of Zollman’s models, but rather Zollman’s models alerted the authors to a potentially salient structural detail of Fermilab’s organization. Similarly, some have argued that cellular automata models are useful to biologists not because they represent any real-world systems, but because they “help us to sensitize our imaginations so that we learn how to notice things we might have missed otherwise” (Weisberg 2013, p. 130). It is difficult to see why, though, absent *any* reason to believe that such models represent relevant features of the world, they would have such a sensitizing effect. If there is no reason to think that Zollman’s model is more likely to predict the performance of epistemic communities than any arbitrary model, how could it be expected to direct our attention in any way?

Finally, a formal model of science could have no value for understanding its target, but nevertheless employ methods that could be used by other models that *are* informative. These models could be methodological pioneers, failing in their intended purpose but nevertheless inspiring further work that does advance our understanding of science and ability to make normative judgments about its social structure. This would make them a genuine contribution to the epistemology of science, but would likely not be satisfying to their creators.

While there are numerous possible aims of formal models of science, the ability to make normative judgments and prescriptions about the social structure of science must remain primary. It is the defining aim of systems-oriented social epistemology, and is what is explicitly claimed of them by their authors.

4 Predictive and representational accuracy

In this section I will propose that formal models of science be evaluated according to their representational and predictive accuracy. Although these models have frequently been discussed on a case-by-case basis, there have been almost no attempt to offer a general account of how to evaluate such models. The account I offer aims to capture the sorts of practices employed by philosophers and scientists undertaking model evaluation in other domains. Special attention will be given to the evaluation of models in economics, as many formal models of science are adapted from economics.

Philosophers of science have devoted significant attention to the roles of models in scientific investigation, and to the relationship between models, theories, and the world. Much of this discussion has been devoted to ontological questions: what *are* models? Various accounts have been proposed, including sets of sentences, mathematical structures, caricatures, and fictions (Frigg and Hartmann 2012). They might exist in the physical world as in the model of the San Francisco bay discussed by Weisberg (2013), in computer code as are many of the complex models employed by climate scientists, or as purely abstract entities such as the model pendulum. They might be instantiations of theories, with no particular real-world target in mind, or they might attempt to capture a single target system, such as a bay, a nation's economy, or the Earth's climactic system.

Because models exist in so many forms to so many purposes, and because there is no consensus among philosophers regarding their ontological status, neither is there any universal account of how they ought to be evaluated. Nevertheless, all philosophical accounts of learning from models appeal to some sense of similarity between models and their target systems as the basis of evaluation (Weisberg 2013, p. 142).

One way in which a model might be similar to its target is structurally: the model might identify certain entities and relationships between those entities, and those entities and relationships might be similar to entities and relationships in the world. Schelling famously modelled cities as regular grids, where each point on the grid is potentially inhabited by a resident of one ethnicity, represented by a penny, or another, represented by a dime. In each iteration of the model, residents assess the ethnic composition of their immediate neighbours, and if the ratio of neighbours of their ethnicity to neighbours of the other ethnicity falls below some threshold, will move to a random location that satisfies their preferences (Schelling 1969). Discussing Schelling's model, Sugden (2008) argues,

> What Schelling has done is to construct a set of imaginary cities, whose workings we can easily understand. In these cities, racial segregation evolves only if people have preferences about the racial mix of their neighbours, but strong segregation evolves even if those preferences are quite mild We gain confidence in such inductive inferences, I suggest, by being able to see the relevant models as instances

> of some category, some of whose instances actually exist in the real world. Thus, we see Schelling's checkerboard cities as possible cities, alongside real cities like New York and Philadelphia We recognize the significance of the similarity between model cities and real cities by accepting that the model world could be real that it describes a state of affairs that is credible, given what we know (or think we know) about the general laws governing events in the real world. (502–503)

Schelling's argument is that racial segregation can evolve in cities even when racial preferences are mild. In Sugden's account, we have reason to believe Schelling's argument because his model cities, constructed out of pennies and dimes on a paper grid, are structurally similar to actual cities. They contain 'people' who behave in similar ways to actual people. These people have neighbours, just as they do in actual cities, and preferences about those neighbours. If they are sufficiently dissatisfied with their current location, they can move to another location that better satisfies their preferences, just as we know people often do. While there are many dissimilarities between Schelling's 'cities' and real cities such as New York or Philadelphia, there are also differences between those real-world cities. Schelling's cities are similar enough to real cities and his coins are similar enough to real people that we can imagine them as real, and believe that people could make living choices similarly to how they do in the model. This similarity is what I mean by representational accuracy. The more similar the model appears to its target—the more representationally accurate it is—the more reason we have to accept its claims.

Representational accuracy is not the only desideratum for scientific models. Indeed, (Morrison 2015) claims that, "perhaps the most important feature of a model is that it contains a certain degree of representational *inaccuracy*" (122, my emphasis). Morrison sees models as vehicles for reasoning about the world, and so a major function of representation is to "conceptual[ize] something in a way that makes it amenable to theoretical or mathematical formulation" (129). This is a common thread in philosophical accounts of modelling: models are tools for reasoning, and so they must necessarily idealize or abstract the world; their usefulness depends crucially upon representational inaccuracy. As a case in point, Morrison relates James Maxwell's ether model of electromagnetic forces. Maxwell described the propagation of electromagnetic waves through space as a mechanical rotation of wheels and vortices. According to Morrison, "Maxwell's ether model was a useful representation, not because it was true or approximately true but because it could be used to generate hypotheses about how electromagnetic waves were propagated through space. But that model needed to represent the ether as a mechanical system if it was to yield any useful results"(128). Maxwell did not attempt to construct a representationally accurate model of the ether for his theoretical explorations—he did not claim that space actually contained wheels and vortices—but rather he constructed a model that allowed him to apply the analytical tools at his disposal for solving the problem.

Nevertheless, Maxwell's model could not be entirely divorced from considerations of representational accuracy. His model was constrained by electromagnetic theory. For example, the flow of electricity, modelled as the rotation of wheels, was constrained by Ampere's law, and Maxwell used Hooke's law to model the elasticity of the ether (Morrison 2015, pp. 103–105). Thus while there were clear and intentional *qualitative* representational inaccuracies in Maxwell's model, he constrained its behaviour according to empirically grounded theories—it was *quantitatively* accurate. Without at least some link to empirical reality, there would have been no reason to suppose that Maxwell's model was anything more than clever fantasy.

Cartwright (2005) agrees with Morrison about the importance of representational inaccuracy. In Cartwright's account, Galilean idealization, the elimination of disturbing causes, is what allows models to identify tendencies or capacities in nature. Models that very accurately represent a specific system might be very good at predicting the behaviour of that system, but will not be generalizable to other, similar, systems. It is Galilean idealization that allows for generalizability. For example, the standard model of a simple pendulum consists of a massive bob connected to a rigid, massless rod, in turn connected to a frictionless pivot. The pendulum oscillates over a small angle and is unaffected by air friction. These idealizations, which strip away disturbing causes such as friction and elasticity, are Galilean. The forces that remain, gravity and the tension of the rod, exist in the target system. In Cartwright's account, these "stripping away" idealizations isolate the causal mechanism responsible for pendulums, in general, to display harmonic motion.

With respect to economic models, however, Cartwright argues that they too often rely on representational inaccuracies that are *not* Galilean.[2] Non-Galilean idealizations do not merely strip away potentially disturbing causes, but introduce new ones. For the pendulum case, a non-Galilean idealization would *add in* a new causal mechanism, not present in actual pendulums—perhaps an electric motor at the pivot point. In economics, a common Galilean idealization is to assume that a person is solely motivated by pecuniary interests—all they care about is money. Real people care about much more than money, but many of us do care about money, and the economist might wish to discover behavioural tendencies that result from that interest. A non-Galilean idealization would be to assume that people are *perfectly rational*: able to instantaneously calculate the expected consequences of every possible decision and act accordingly. While both assumptions are false, the first identifies a feature of real people while the second attributes to us something that no actual person possesses.

Theoretical economics is unable to identify genuine social tendencies, Cartwright argues, because of its reliance on Non-Galilean idealization. Economics possesses only "very meagre" theoretical generalizations from which to generate economic models, and those generalizations (such as that people are motivated

[2] Philosophers often distinguish between Aristoltean idealization, which ignores irrelevant features of a system, and Galilean idealization, which deliberately distorts the system. According to this taxonomy, Cartwright's "Non-Galilean" idealizations would actually be a subset of Galilean idealization. For simplicity I maintain Cartwright's terminology.

by money, or that demand for a product is inversely related to its price) are incapable of generating many interesting results (125). Instead, the results of economic models are generated by non-Galilean idealizations that are unconstrained by economic theory. She describes, for instance, how Robert Lucas's model demonstrating the neutrality of money details pages and pages of assumptions including that all trade is confined to two separate markets, that there is a random distribution of people into "young" and "old", that no communication between markets is possible, that within each market there is a single market clearing price, and that the money supply is known by all participants (126). It is these assumptions, Cartwright argues, that generate Lucas's conclusions, not economic theory. Because their results depend not on Galilean idealization constrained by theory but non-Galilean idealization unconstrained by theory, Lucas's and other models in theoretical economics, she claims, are incapable of identifying genuine tendencies in economies.

Note the contrast between Lucas's model and Schelling's model of segregation discussed above. While Schelling's model involves much Galilean idealization—people *only* care about the ethnicities of their neighbours—its results do not crucially depend on non-Galilean idealization.[3] Schelling invites his readers to experiment with varying rules for how "tolerant" people are of living as ethnic minorities, and the results of his model—that cities spontaneously segregate into large patches of ethnically homogeneous neighbourhoods—remain, as long as people have *some* preference regarding the ethnicity of their neighbours. Since people do seem to have such preferences, this amounts to a Galilean stripping away of interfering causes, not a non-Galilean introduction of new causes. Further, Schelling's model contains none of the common non-theoretical assumptions of mathematical economics—perfect rationality, single prices, and so on. Schelling's model is mathematically unsophisticated, but this is a virtue, as we can be confident that its dynamics are not the result of empirical assumptions independent of economic theory.

We can contrast Cartwright's account of idealization in economics as exemplified by Lucas with Morrison's account of idealization in physics as exemplified by Maxwell. In both cases Galilean idealizations—representational inaccuracies—are involved, but with differing consequences. The crucial difference is that Maxwell's idealizations and his subsequent reasoning are constrained by theories or empirical regularities, such as Ampere's Law and Hooke's Law. The results of his model-based reasoning are credible precisely because they are constrained in this way. Lucas's model, by contrast, has no such constraints. Instead his results are due to non-Galilean idealizations adopted for analytical tractability. It is this qualitative difference in representational accuracy that makes all the difference, in Cartwright's account, for how we judge the credibility of models.

There are two lessons I want to draw from this discussion of differing kinds of representational inaccuracy. First, representational accuracy does matter. We

[3] This is not to say that Schelling's model involves no non-Galilean idealizations. For instance, it assumes that moving is costless and that individuals of either colour are equally free to move to any vacant location.

can justifiably make inferences about how the world is likely to be from models based on judgments about how accurately those models represent the part of the world they seek to describe. Knowing nothing else about either model, there is more reason to expect that Maxwell's predictions about the behaviour of electromagnetism will be correct than to believe that Lucas's predictions about the operation of money will be correct, just because Maxwell's model is more constrained by empirically-informed theory than Lucas's. Second, representational accuracy is not a single, quantitative scale. While in some domains quantitative comparisons of representational accuracy might be possible, qualitative assessment of the *kind* of representational inaccuracy displayed by models is what matters. Not all inaccuracies are created equal.

While assessing representational accuracy requires some comparison between the structure of a model and the structure of its target, predictive accuracy involves comparing predictions generated by the model with observations of its target. Strictly speaking, predictive accuracy is not a measure of the model itself, as models do not make predictions about the world, but a measure of the predictions modellers make *based* on the behaviours of models. Schelling's model, for example, is incapable of making predictions itself; it is a non-linguistic entity composed of pennies and dimes. But based on observations of the model, one can make predictions about its intended target. For example, the boundaries between segregated neighbourhoods in Schelling's model tend to shift through time. We might assess the predictive accuracy of his model by checking whether segregated neighbourhoods in actual cities display the same boundary-shifting behaviour. If they do, this could increase the credibility of its other predictions, such as that a new, ethnically-diverse, city will likely become increasingly segregated if its residents have even mild preferences regarding the ethnicities of their neighbours and they are free to choose where they live.

Milton Friedman famously argued that predictive accuracy is *all* that matters for assessing economic theories. He argues:

> the relevant question to ask about the "assumptions" of a theory is not whether they are descriptively "realistic," for they never are, but whether they are sufficiently good approximations for the purpose in hand. And this question can be answered only by seeing whether the theory works, which means whether it yields sufficiently accurate predictions. The two supposedly independent tests thus reduce to one test. (Friedman 2008, p. 153)

Similarly to Morrison and Cartwright's claims about models, Friedman argues that in order to be useful theories must "abstract the common and crucial elements from the mass of complex and detailed phenomena to be explained" (153). Representational inaccuracy is a virtue. But whereas for Morrison and Cartwright the details of that idealization matter for assessing the model, for Friedman it is only predictive success that matters. He does not dispute that there must be some relationship between the structure of a theory (or model) and the world, but argues that no *assessment* of the representational accuracy can determine whether the theory (or model) should be accepted. All that can be usefully assessed is whether it "yields sufficiently accurate predictions" for its intended purpose.

Friedman's narrow instrumentalism—that all that matters is predictive success for a narrow range of phenomena—has been widely criticized by philosophers. One criticism is that many of the "predictions" in economics that Friedman describes cannot actually be assessed. For instance, Friedman argues that economic theory predicts that firms will maximize profits, and if this prediction is successful we should not care about whether the people running these firms really are perfectly rational. But how could we assess, even in principle, whether firms really do maximize profits (Simon 2008)? Another criticism is that, without assessing a model or theory's representational accuracy, it is impossible to know how to infer from past predictive success to future success in new domains. (Hausman 2008) compares theory assessment to test driving a car. Looking under the hood, Hausman argues, is clearly a worthwhile activity in assessing the car's likely future performance, especially when it is difficult to test the car in a wide variety of circumstances. Similarly, examining the details of a model or theory's representation can be useful for assessing whether it is likely to succeed for predicting phenomena beyond those for which it has already been successful. The virtue that Friedman sees in idealization—that it allows for generalization by abstracting the "common and crucial elements" of economic circumstances—is defeated by his restriction of theory assessment to narrow predictive success. Generalization requires, in Hausman's account, some assessment of representational accuracy.

It is worth asking whether there is any principled distinction between representational and predictive accuracy. For example, does comparing the pattern of segregation in Schelling's model cities with segregation in actual cities constitute an assessment of representational accuracy (there is a similarity in the structure of model and target) or predictive accuracy (from the behaviour of the model we predict that cities whose residents display even mild racial preferences will evolve segregated neighbourhoods)? I believe a useful distinction can be made, along at least two axes. First, epistemically: assessing representational accuracy involves comparing what we already know about a system with the model while assessing predictive accuracy involves comparing novel behaviours of the model with the behaviour of its target. Second, ontologically: assessing representational accuracy involves assessing the structural features of a model (what entities and relationships constitute the model) while assessing predictive accuracy involved assessing behaviour (what results from those entities and relationships). It is not the final configuration of residents in Schelling's model that is predictive, but the evolution of diverse neighbourhoods into segregated ones.[4] Therefore although there might be ambiguity in categorizing some particular

[4] As a reviewer observed, this is a weak, qualitative prediction. Schelling's model, like some formal models of science, relies on the robustness of this behaviour for its explanatory appeal. As discussed below, what distinguishes Schelling's model from most formal models of science is not its predictive accuracy, but that this behaviour is generated by Galilean (stripping away) idealizations.

assessment activity, there does seem to be a meaningful distinction between the two.

Although distinct, representational and predictive accuracy do not operate in isolation. The link from representational to predictive accuracy seems clear. The representational accuracy of Schelling's model, so far as it goes, is evidence that predictions derived from it will be correct. That is not to say that the relationship is perfectly monotonic; models may become less predictively accurate due to increases in representational accuracy or more accurate due to decreases. For instance, Winsberg (2006) describes the insertion of "artificial viscosity" into fluid dynamics models to increase predictive accuracy. Since artificial viscosity introduces a property into models that is known not to exist, it decreases their representational accuracy. But in doing so it increases predictive accuracy by compensating for the inability of computer simulations to properly model shock waves. Conversely, removing this false feature from fluid dynamics models would increase representational accuracy but decrease predictive accuracy. Exceptions such as this aside, however, the reason that modellers strive for increased representational accuracy at all is that they believe that by doing so they will be able to better predict the behaviours of their target system. This is the driving intuition behind, for example, the development of global climate models, which over the last decades have gradually added more and more relevant features of the climate system and increased the resolution of their representation of existing features. As discussed below, these improvements have in turn led to improvements in predictive accuracy. If we could not usually infer from increased representational accuracy to increased predictive accuracy, it is difficult to see why scientists would pursue modelling at all.

Inferring from increased predictive accuracy to increased representational accuracy is less obvious. Regarding scientific *theories*, the no-miracles argument claims that the only or best explanation for predictive success is representational success: truth. However, there seem to be clear instances of increased predictive accuracy *not* implying increased representational accuracy. For example, adding epicycles to Ptolemy's Earth-centred model of the universe increased its ability to predict the location of the sun and planets in the sky, but did not indicate increased representational accuracy (Kuhn 1985). If anything, it signalled a decrease. Friedman can perhaps be read as denying this implication in general. On the other hand, even Friedman suggests that one can infer some degree of representational accuracy from predictive success. For instance, in the above-quoted passage, he implies that the only test of whether the assumptions of a theory are "sufficiently good approximations" of reality is whether it "yields sufficiently accurate predictions" (Friedman 2008, p. 153). That is, he suggests that you can infer from accuracy of prediction to some measure of goodness of approximation—representational accuracy. Friedman does not deny a link between predictive accuracy and representational accuracy; he denies that there is any way of *assessing* representational accuracy independently of predictive accuracy.

Indeed, from a Bayesian perspective, if increased representational accuracy tends to imply increased predictive accuracy, increased predictive accuracy must tend to imply increased representational accuracy. Taking $p(PS)$ as the probability that the model will generate a successful prediction and $p(RS)$ as the probability that the model will pass some test of its representational accuracy, then by Bayes' law:

$$p(PS|RS) = \frac{p(RS|PS)p(PS)}{p(RS)}$$

If $p(PS|RS) > p(PS)$—if the probability of a model generating a successful prediction given that it has passed a test of its representational accuracy is greater than it is before passing the test—then it must be the case that $p(RS|PS) > p(RS)$. It must be true that the probability of passing a test of representational accuracy is higher if it has made a successful prediction than if it has no, all else being equal. So while increased predictive accuracy does not always imply increased representational accuracy, there must be a tendency to do so as long as there is a tendency for increased representational accuracy to imply increased predictive accuracy.

Sometimes predictive accuracy cannot be assessed, or at least not in a way to ground sufficient confidence in a model's further predictions or representational accuracy. One way of ameliorating this difficulty is through robustness analysis. Robustness is a measure of whether and to what degree a model's predictions rely on the details of its representation—either in terms of its qualitative structure or quantitative parameters (Weisberg 2013, pp. 162–166). The more robust a model's predictions are to either sort of change, the more reason there is to believe that its predictions are true, even if those predictions are impossible to verify directly.

Orzack and Sober (1993) argue that robustness alone can have "heuristic" value but little else; it cannot be used as a substitute for predictive accuracy in assessing a model. On their account, models assessed only on the basis of robustness can be useful as engines of inquiry, but little else. They imagine robustness as follows. Suppose there is a set of models $M_1..M_n$, perhaps generated by varying the parameters and structure of a model as described above. This set of models robustly predict some property R if all members of the set display R. If one of these models is known to be "true" (perfectly representationally accurate), and all of these models imply R, then R must be the case (538). Outside of this special case, unlikely to ever be realized by a set of existing models, they argue that robustness cannot confirm properties such as R without empirical data. One possibility they consider is expanding the set of models to "all possible models" and sampling from these models at random (539). Perhaps if all or most of these models exhibit R then it is a robust property likely to be true in the target system as well. However, they judge the notion of "all possible models" incoherent and do not see a method of making such random draws. Therefore there is no determinate way of adjusting our belief in R as a genuine property of the target system based on robustness analysis alone.

Weisberg (2006) disagrees, arguing that robustness can be confirmatory even in the absence of direct empirical data. Slightly simplifying Weisberg's account, again consider a set of models $M_1..M_n$ of a target system that all exhibit property R. Further, suppose that it is possible to identify some common structure, S, that is instantiated by each model and is identified as generating R. If the target system also possessed S, then it too would exhibit R, absent any interfering factors, but it would require (possibly unavailable) data to assess whether it does possess S. Instead, Weisberg argues that if $M_1..M_n$ are "sufficiently heterogeneous"—that is, vary enough in their parameters, structures, and modes—then there is reason to believe that the target system too will possess this structure and therefore exhibit R (again, assuming no interfering factors). This procedure, Weisberg argues, is not nonempirical, because the scientists who generated $M_1..M_n$ used structures that had been previously demonstrated to make correct predictions, what he calls "low-level confirmation" (740). For example, the Volterra predator-prey model is based on coupled differential equations, which themselves have been empirically confirmed (741). It is this low-level confirmation that allows for the inference that the properties of a heterogenous set of models of a target system are relevant to inferring properties of the target system. Otherwise there would be no reason to suppose $M_1..M_n$ have anything to do with their target. Therefore robustness analysis can be seen, in Weisberg's account, as a way of exploring the consequences of *already established* empirical regularities in different contexts. This point will be central to my discussion of robustness in formal models of science.

There is a clear connection between Weisberg's account of robustness and Cartwright's account of Galilean idealization. In both cases what matters is that the conclusions drawn from a model or set of models about a target system are connected to it through empirical investigation. In Cartwright's account, the problem with non-Galilean idealization is that the model behaviours are not due to empirically-informed theory but to assumptions adopted purely for the purposes of making a problem analytically tractable. In Weisberg's account robust properties of a family of models are likely also present in the target system because those properties are ultimately the result of empirically-confirmed principles. Models in both these accounts are a vehicle for discovering previously unrealized consequences of our knowledge of the world. Models do not generate knowledge about systems from nothing.

Contrary to these accounts, Grüne-Yanoff (2009) argues that it is possible to learn from models that "are assumed to lack any similarity, isomorphism or resemblance relation to the world to be unconstrained by natural laws or structural identity, and not to isolate any real factors", what he calls "minimal models" (83). Following Lenker (1999), Grne-Yanoff argues that minimal models are judged based on their "credibility" or "plausibility" rather than any direct correspondence to their targets. These models inhabit parallel worlds that represent how our world *could be* (95). We can learn from such models that certain propositions thought to be *impossible* are instead *possible* because they are realized in a parallel world by a credible model (97). Grne-Yanoff's account seems well suited to the engines-of-inquiry function of models, as it gives reason for investigating features suggested by the models in target systems. If models credibly suggest that a property we believed *impossible* of a system is possible, it could be worthwhile to look for the realization of that possibility using other means. However, as Grne-Yanoff is careful to point out, by his account such models are not informative regarding properties we already believed possible. For example, nobody believes it *impossible* that segregated cities could exist when residents have only mild racial preferences. So unless Schelling's model has some recognized similarity with real cities—unless there is reason to believe it has some degree of representational accuracy—it would be, under Grne-Yanoff's account, uninformative.

I this section I have argued that models can be assessed according to two broad criteria: representational accuracy and predictive accuracy. These criteria are distinct, but related. Regardless of which criterion is being used to assess a scientific model, however, the model's success must be connected to empirical data. Morrison argues that models in theoretical economics fail because their predictions are derived from non-Galilean idealizations rather than from idealizations of empirically-derived theory, while Weisberg argues that robustness is informative when models are based on lower-level empirical confirmation. Assessing predictive accuracy, meanwhile, directly involves comparison between predictions derived from models and empirical observation.

5 Assessing models in science

This section will demonstrate how evaluating models according to their predictive and representational accuracy can be applied to familiar models in science. These models have also been selected to allow for productive comparison with formal models of science.

The first model, I wish to consider is the Global Climate Model (GCM). GCMs are models that attempt to simulate the entire climactic system under different natural and human scenarios. They are instructive for two main reasons: they demonstrate how confidence in a model's predictive power can come from both assessments of representational and predictive accuracy, and they show the prospects of robustness analysis to increase confidence in the predictions of a set of related models. They also demonstrate just how difficult it is to model complex phenomena sufficiently well to drive policy changes—the ultimate aim of formal models of science.

Scientists make great effort to maximize the representational accuracy of GCMs. At their core, they simulate the atmosphere and oceans based on fundamental physical theory, such as fluid mechanics and thermodynamics (Edwards 2010, p. 145). Models contain a representation of the Earth's surface, oceans, and atmosphere divided into grids horizontally and vertically, and slices vertically. Although debate exists within the modelling community over whether increasing the "realism" of models should be an end in itself (rather than merely a means to increasing predictive power) (Edwards 2010, p. 345), models have steadily increased in their complexity and ability to accurately represent the climate system (Abiodun et al. 2013). Corresponding to this trend is an increase in models' ability to predict (or retrodict) important quantities such as surface temperature and precipitation. This correspondence is not monotonic, as "every bit of added complexity, while intended to improve some aspect of simulated climate, also introduces new sources of possible error... and new interactions between model components that may, if only temporarily, degrade a model's simulation of other aspects of the climate system"(824). Nevertheless, this trend gives support to the claim that increasing representational accuracy can also be expected to increase predictive accuracy.

Robustness of predictions across climate models is frequently cited as a reason to trust those predictions. For instance, (Lloyd 2009) argues, following Weisberg's account of robustness, that because a wide range of climate models simulate a twentieth century warming trend and they all attribute this warming to greenhouse gasses, there is reason to believe that greenhouse gasses are in fact largely responsible for twentieth century warming (Lloyd 2009, p. 220). However, there is considerable disagreement about what can justifiably be inferred from such robust predictions. Katzav (2014) argues, for instance, that because climate models incorporate many known-to-be-false assumptions that are key to their predictive success, they cannot be confirmed and instead can only express ranges of possibilities. That the epistemic role of robustness is so controversial even in this near-ideal case—where the main model dynamics are based on well-understood physical theory, there is a concrete target, and models are made as representationally accurate as possible—should give philosophers of science pause in appealing to robustness in their own models.

The next model worth discussing in Schelling's model of racial segregation (Schelling 1969), introduced above. Schelling's model is perhaps a much more fair basis of comparison for formal models of science than global climate models. Formal models of science, after all, do not aim to make quantitative predictions, while quantitative predictions are of central importance for GCMs. Schelling's model, like models of science, is highly idealized and qualitative. Its credibility results not from the quantitative predictive success of any particular instantiation, but from robustness: the same segregated outcomes result from a wide variety of initial conditions and agent preferences. Representational accuracy is not established through empirical observation, but through plausibility: his agents behave in ways that people might plausibly decide about where to live and his grid is a plausible representation of a suburban landscape. Following Cartwright, the reason that Schelling's model credibly identifies a genuine tendency in real-world cities is that its idealizations are Galilean: his model strips away interfering causes that operate in our world, but the causes that do operate in his world also operate in ours. This is an advantage of Schelling's model over most formal models of science, whose idealizations are not Galilean by Cartwright's account.

Nevertheless, with robustness acting as a proxy for predictive accuracy and plausibility standing in for any rigorous account of representational accuracy, Schelling's results deserve skepticism. His model allows us to avoid the uncomfortable conclusion that highly segregated U.S. cites are the result of significantly racist attitudes. Since for this reason we may be predisposed to accept Schelling's argument, there is even more reason to be cautious of his results absent further evidence. However, because the components of his model have observable analogues, it is amenable to empirical confirmation. One can survey racial preferences, map neighborhoods by ethnicity, and watch how they evolve. Indeed, social scientists have done such studies, citing Schelling's model as an impetus for their studies (Massey and Denton 1993, p. 96). While they have confirmed the essential dynamics of his model, there are important caveats: black people tend to be far more willing to live in majority-white neighborhoods than white people, and black people often face systemic barriers that make moving into majority white neighborhoods difficult. These results show that white racism plays a more significant role in U.S. segregation than Schelling's model suggests. Nevertheless, because Schelling's idealizations were largely Galilean, the cascade mechanism driving his results identified a genuine tendency in actual cities.

Finally, consider the example of Arrow and Debreu's model of general equilibrium, the paradigm achievement for mathematical microeconomics (Arrow and Debreu 1954). Arrow and Debreu's model is both a methodological predecessor of economic models of science such as Kitcher (1990) and Strevens (2003), and a potential aspirational example for current formal models of science. It is, after all, highly technically sophisticated, highly cited, and earned both of its authors Nobel Prizes. It pioneered mathematical techniques central to modern microeconomics and spawned an industry of models exploring the consequences of relaxing its assumptions and applying its concepts to new domains (Geanakoplos 1998). As formal models of science hope to, it offered a normative judgment about its target system, the market. Finally, just as for most formal models of science, it is not amenable to empirical testing and is instead motivated by plausibility and realism. However, it is also emblematic of an approach to economics that many view as a wrong turn, against which microeconomics has only recently started to recover with a renewed focus on empirical study (Angrist et al. 2017). Despite all of its disciplinary achievements, it is not clear that the general equilibrium model or any of its methodological successors are genuinely informative about the properties of real markets.

The Arrow-Debreu economy consists of commodities, firms, and consumers. Consumers sell their labor to firms in exchange for commodities. Consumers have preferences over possible consumption plans while firms seek to maximize profits. The model's assumptions include that commodities are infinitely divisible, consumers and firms are perfectly rational, consumers are never fully satiated, present and future prices are known to all, and each individual accepts those prices as given. These assumptions are, of course, false for any real economy, but nevertheless Arrow and Debreu argue that an advantage of their model is that its assumptions are "weaker and closer to economic reality" that previous general equilibrium models (Arrow and Debreu 1954, p. 266). In some sense the model is a plausible representation of a real economy: consumers, producers, and commodities have clear analogues in real economies and they stand in similar relations in both the model and the world. The assumptions of the model, while false, are arguably reasonable idealizations.

However, just as for of Lucas' model discussed above, Arrow and Debreu's assumptions are non-Galilean. They do not just ignore the possibility that firms might not always seek to maximize profits, but impute properties to consumers, firms, and commodities that are *impossible* for consumers, firms, and commodities to possess in the real world. Thus the dynamical structure of real markets, with quasi rational, imperfectly informed, learning individuals variously competing and cooperating, is entirely absent from the general equilibrium model. The causal structure of the model is qualitatively different from the causal structure of its target, and this is what really matters for assessing representational accuracy. As Hausman (1992) puts it,

> If, as seems likely to me, there are systematic and important failings of human rationality, and economic behaviour is significantly influenced by many motive forces, apart from consumerism and diminishing marginal rates of substitution, then equilibrium theory is not a very good theory, whether or not there is anything better. If it leaves out important causes, then no mathematical expertise or elegance in modeling will make equilibrium theory into a good theory. (280)

While the Arrow-Debreu model itself says nothing about the robustness of its results, many successors have explored the consequences of relaxing one or another of its assumptions. Grossman and Stiglitz (1980), for instance, shows that when information is costly, prices can approach but never reach their equilibrium values. Relaxing other assumptions, such as constant returns to scale for firms, is not nearly so benign (Geanakoplos 1998, p. 121). While I am not aware of any systematic attempt at a robustness analysis in the mode of Weisberg (2006), the general approach of Arrow and Debreu's successors is reminiscent of the underlying idea: by exploring how models with different assumptions behave, economists might be able to learn what features of the Arrow-Debreu economy exist in real economies and in what contexts. However, this is a very limited sort of robustness, and even to the extent that there are robust properties of general equilibrium models, they do not meet Weisberg's requirement that there be low-level confirmation of the models' component parts.

Arrow and Debreu's chief achievement was to show that, given their assumptions, there always exists a set of prices such that the quantity of commodities produced by firms exactly equals the quantity of commodities demanded by consumers. No commodity goes unconsumed and no consumer demands a commodity that is not produced. The normative implication of this result is that, for a competitive economy conforming to Arrow and Debreu's assumptions, every distribution of commodities is Pareto efficient. Conversely, every Pareto efficient distribution is achievable by an Arrow-Debreu economy given some initial distribution of resources. A Pareto efficient distribution is one where no individual's preferences could be better satisfied without making another individual less satisfied. This link between perfect competition and Pareto efficiency is often invoked as a justification for economic policies intended to move actual markets closer to Arrow and Debreu's ideal Hausman and McPherson (2008).

Because, all else being equal, Pareto efficiency is better than Pareto inefficiency, and ideal markets are Pareto efficient, one might conclude that making markets more ideal will make people better off. However, since Arrow and Debreu's model is not robust to arbitrary de-idealizations, any particular de-idealization "may well *diminish* rather than improve efficiency" (Hausman and McPherson 2008, p. 242). So invoking Pareto efficiency as a justification for market reforms on the basis of Arrow and Debreu's model is unjustified.

So while Arrow and Debreu's general equilibrium model may seem a worthy one for formal models of science to emulate, given its undeniable disciplinary success, upon further examination its status is much less clear. What do we learn about markets from examining it? What normative lessons can be drawn? Perhaps nothing and none.

6 Evaluating formal models of science

The primary aim of this section is not to criticize particular models of scientific organization. Such critiques have already been made. Among others, Muldoon and Weisberg (2010) examines the models of Kitcher (1990) and Strevens (2003); A. Thompson (2014) devastatingly appraises Hong and Page (2004); both Thoma (2015a) and Alexander, Himmelreich, and C. Thompson (2015a) question the results of (Weisberg and Muldoon 2009); and Rosenstock, Bruner, and OConnor (2017) critiques (K. Zollman 2007) and similar models. Instead, the aim of this section is to offer a general assessment of the prospects of formal models of science. While critiques of individual models are abundant, there has been very little general discussion of the practice of modeling the social structure of science.

As far as I am aware, the one[Michael Thicke 2018–08–29, 3:16 PM

I discuss 2?] published systematic review is (Martini and Pinto 2017), which argues that to make progress formal models of science need to connect with empirical data. I am generally sympathetic to their argument. What I hope my discussion adds to theirs is a theoretical understanding of what is going wrong with formal models of science which can serve as a basis for a wider methodological conversation about the prospects of formal models of science to more credibly asess the behaviour of scientific communities. It should also help to understand the implications of more targeted examinations. For example, many of the above-mentioned critiques question the robustness of their examined models. But what sorts of robustness matter, and what can we infer about science from models that do exhibit robust features of one sort or another?

The subject matter of formal models of science, knowledge, makes evaluation based on predictive accuracy inherently difficult. Whereas GCMs can compare their predictions with observed temperature, precipitation, and other climate variables, it is much harder to compare the speed and accuracy of a scientific consensus with that predicted by K. Zollman (2007). However, relevant comparisons are not impossible. Dodge (1996), for instance, evaluates consensuses such as over continental drift in the mid-twentieth century and the "central dogma" of molecular biology—that DNA is the bearer of inherited traits—around the same period based on their subsequent empirical success. A consensus is justified, according to Solomon, if the accepted theory has all of the empirical successes and the rejected theory has none. While Solomon's account of scientists choosing theories based on empirical and non-empirical decision vectors is at most an "informal" model of science, she is able to demonstrate its usefulness by applying it to concrete cases. Indeed, Solomon's detailed historical case studies highlight the complete lack of similar studies in the formal modelling literature. Formal models of science could go a long way toward increasing their credibility by demonstrating such applications, especially if they are able to show that their models predict collective behaviours.

Martini and Pinto (2017) suggest citations as one potential source of data for formal models of science. Citations have the potential to function as proxy measures of many salient features of science. For instance, (Perović et al. 2016) use citations to evaluate the success of particle physics experiments, and so use citation counts as a proxy measure of epistemic significance. Hicks (2016) uses between-discipline citations to measure the "social validation" of research results. The distribution of citations between competing researchers could also be used to evaluate the winner-takes-all model of Strevens (2003). However, as Martini and Pinto (2017) argue, most current formal models do not make predictions that are amenable to citation analysis, and it is unclear whether a model such as that of Weisberg and Muldoon could even in principle be adapted to be compared with citations, or any other empirical data (233).

Because directly evaluating the predictive accuracy of formal models of science is difficult-to-impossible, robustness is frequently invoked as an alternative. Muldoon and Weisberg (2010) criticize the models of Kitcher (1990) and Strevens (2003) for not being robust to changes in scientists' ability to view each other's decisions. Alexander, Himmelreich, and C. Thompson (2015b) criticizes Weisberg and Muldoon (2009) for not being robust to changes in agent behaviour. K. Zollman (2007) claims his model "suggests that there is a *robust* tradeoff between speed and reliability"(575, my emphasis) in scientific communication networks, while Rosenstock, Bruner, and OConnor (2017) argues that, on the contrary, Zollman's results are not robust to small changes in their parameters.

Evaluating models based on robustness assumes that, if their results were robust, this would be reason to give them credence as modelling real features of scientific communities. Rosenstock, Bruner, and OConnor (2017) argue, for instance, that the benefits of epistemic diversity suggested by (K. J. Zollman 2010) holds robustly in his and similar models, and therefore is more likely to be true of science than his claim that sparsely-connected communication networks can be more reliable than well-connected ones. They claim:

> Highly robust phenomena are more likely to have applicability to the real world because the conditions under which they occur are more likely to hold. When epistemic network effects are highly robust, it makes sense to take them more seriously as important findings for

> real world communities. (251)

Rosenstock, Bruner, and OConnor (2017) acknowledge that the epistemic significance of robustness is controversial, but they do not do justice to the details of philosophical accounts. They cite Weisberg (2006) in favour of robustness, but as discussed above, Weisberg does not defend robustness unequivocally. Robustness is only epistemically significant, in his account, when there is lower-level empirical confirmation of the models, and those models are sufficiently heterogeneous. There is no such empirical confirmation of formal models of science: the behaviours and dynamics of their modelled scientific communities have not been shown to correspond to those of real communities through any sort of empirical comparison. Showing that essentially free-floating models have robust properties should not be seen as increasing the likelihood that those properties also exist for real scientific communities. Robustness might be able to establish *possibility*, as (Grüne-Yanoff 2009) argues it does in the case of Schelling's segregation model, but one might then ask what properties of science were thought to be impossible before the existence of formal models?

While robustness is often cited as a proxy for empirical assessment of predictive accuracy, realism or plausibility is used as a proxy for empirical assessment of representational accuracy. For example, Thoma (2015a) argues that because her agents behave more realistically than those of Weisberg and Muldoon, her claims about the benefits of diversity in science are more credible. One might hope that plausibility and robustness can work together to lend credence to a model's predictions. Plausibility, perhaps, can establish the sort of low-level confirmation demanded by Weisberg, and robustness in turn can discover properties likely to exist in real scientific communities. I am skeptical, though, that such an approach can be defended.

There are two good reasons for skepticism. First, the plausibility established by most formal models of science is very weak; while there might be *some* similarities between the organization of scientific communities and the structure of these models, it is often a very distant sort of similarity. For instance, Hong and Page (2004) model individuals as triplets of integers that correspond to search strategies of a number line, and measure diversity according to the differences between triplets. Perhaps their model says something significant about searching number lines (though A. Thompson (2014) argues they fail at even this), but what licenses their claim that this says anything about diverse individuals solving arbitrary problems? The internal dynamics of their model are highly abstract: the triplets specify simple algorithms that do not bear any clear resemblance to how scientists or any individuals might actually make methodological choices.

Along similar lines, Thicke (2017) argues that Muldoon and Weisberg's epistemic landscape seems to have little to do with actual scientific practice. The z-axis of their epistemic landscape corresponds to the "epistemic significance" of a particular approach to solving some problem of scientific interest, but McKenzie Alexander wonders whether it makes sense to assign an objective value to such a quantity. Even if it does, is it reasonable to suppose that scientists are able to *observe* this quantity, both for their current approach and nearby approaches, as the model requires? If the way that actual scientists make methodological choices bears no resemblance to that used in the model, and the model's dynamics crucially depend on that method of choice, then how can the model credibly represent scientific practice? If the epistemic landscape is incapable of representing actual scientific practice, then no amount of emendation or robustness testing can inform us about the models' intended target.

Regarding formal models of science more generally, McKenzie Alexander argues that their problem stems from the central role of Galilean idealizations to their dynamics. He argues that if a model's dynamics fundamentally depend on such a deliberate distortion, there is no reason to expect that the target, operating according to entirely different dynamics, will exhibit similar properties. Cartwright, discussing models in economics, puts the point somewhat differently, arguing that Galilean idealization is acceptable if it involves only the stripping away of relevant causes, not the invention of new ones, but the intuition is the same: if a model's behaviour is determined by false or empirically uninformed assumptions, there is little reason to expect that predictions derived from it will hold true.

Cartwright might further diagnose the problem with formal models of science as one of even more meagre underlying theories than exist in economics. While economists have established at least some empirical regularities in market economies, what sociological theories of scientific behaviour constrain formal models of science? This diagnosis is similar to Reijula and Kuorikoski's claim that many models of science rely on arbitrary rather than empirically-informed ones.

In contrast to the models discussed so far, Felipe Romero's model of social correction in science does credibly establish its representational accuracy (Romero 2016). In Romero's model, scientists attempt to measure the value of some quantity, such as the degree of correlation between political party affiliation and attitudes towards environmentalism. Sequentially, each scientist performs an experiment to estimate the quantity's value, and aggregates their experimental data with that reported by previous experiments. Romero parameterizes his model according to standard statistical practice in social psychology, using sample sizes and statistical powers similar to those used by scientists. He then simulates the performance of scientific communities under different institutional structures, such as budget constraints reducing sample sizes, only statistically significant results being published, or scientists only reporting positive results. The dynamical structure of his model—the sequence of scientists performing experiments and aggregating their results with previous experiments—is a simplification of actual practice, but qualitatively similar to that performed by actual scientists. The causes driving the dynamics of his model are causes that plausibly drive actual scientific practice. Therefore his robust result—that except under ideal conditions scientists will often fail to self-correct—is epistemically significant. Because Romero first credibly establishes the representational accuracy of his model, his subsequent robustness analysis is informative, and his normative recommendations are credible. It is that first essential step that other formal models of science fail to take.

7 Conclusion

I began this paper with a quotation from Philip Kitcher: “How do we best design social institutions for the advancement of learning? The philosophers have ignored the social structure of science. The point, however, is to change it” (Kitcher 1990, p. 23). The proceeding discussion was based on two premises: that formal models of science are intended to support such normative interventions, and that formal models of science ought to be evaluated according to the same standards philosophers set for models in other disciplines. The first premise is supported by the explicit claims made of the models by their authors. The second, I hope, is uncontroversial. If not, modellers need to explain why not, and offer an alternative method for evaluation.

I have argued that models can be evaluated according to both their predictive and representational accuracy. While robustness and plausibility can to some degree establish each, they can do so only weakly absent empirical comparison. Therefore most formal models of science are incapable of supporting normative conclusions, and instead can only make possibility claims, or act as “engines of inquiry” for further, empirical, investigations.

Formal modellers face a choice: accept this limited role for their models, or seek to establish representational and predictive accuracy in a more rigorous manner. In particular, models can be made more credible by better capturing the dynamic structure of scientific inquiry, by making predictions that are amenable to empirical comparison (such as citation analysis), and parameterizing models according to actual scientific practice.

For consumers of formal models—philosophers, scientists, or policymakers who look to these models for insight into the operation of science—I can only suggest caution. In my view the current generation of formal models, with very few exceptions, is unable to support *any* normative conclusions about science, and should not be invoked in a policy context.

Finally, it is my hope that this paper can help to foster a general methodological discussion about how formal models are constructed and evaluated. Whether or not my account of how to evaluate formal models is accepted, modellers are beholden to present *some* method for evaluating their claims. Without such an account, and a serious defence of their methodology, I see no reason to take these models seriously.

References

Abiodun, Babatunde et al. (2013). “Evaluation of Climate Models”. In: *Climate Change 2013: The Physical Science Basis*. Cambridge: Cambridge University Press, pp. 1–126.

Alexander, Jason McKenzie, Johannes Himmelreich, and Christopher Thompson (2015a). “Epistemic Landscapes, Optimal Search, and the Division of Cognitive Labor”. In: *Philosophy of Science* 82.3, pp. 424–453.

— (2015b). “Epistemic landscapes, optimal search, and the division of cognitive labor”. In: *Philosophy of Science* 82.3, pp. 424–453.

Angrist, Joshua et al. (2017). “Economic Research Evolves: Fields and Styles”. In: *The American Economic Review* 107.5, pp. 293–297.

Arrow, Kenneth and G Debreu (1954). “Existence of an equilibrium for a competitive economy”. In: *Econometrica: Journal of the Econometric Society* 22.3, pp. 265–290.

Cartwright, Nancy (2005). “The vanity of rigour in economics: theoretical models and Galilean experiments”. In: *The Experiment in the History of Economics*. Ed. by Philippe Fontaine and Robert Leonard. Routledge, pp. 118–134.

Dodge, Pryor (1996). *The bicycle*. Paris: Flammarion.

Edwards, Paul N (2010). *A Vast Machine: Computer Models, Climate Data, and The Politics of Global Warming*. MIT Press.

Friedman, Milton (2008). “The Methodology of Positive Economics”. In: *The Philosophy of Economics: An Anthology,Third Edition*. Ed. by Daniel M Hausman. Cambridge University Press, pp. 145–178.

Frigg, Roman and Stephan Hartmann (2012). “Models in Science”. In: *Stanford Encyclopedia of Philosophy*, pp. 1–24.

Geanakoplos, J (1998). “Arrow-Debreu model of general equilibrium”. In: *The new Palgrave. A Dictionary of economics* 1, pp. 116–124.

Goldman, Alvin I (2011). “A Guide to Social Epistemology”. In: *Social Epistemology: Essential Readings*. Ed. by Alvin I Goldman and Dennis Whitcomb. Oxford University Press.

Grossman, SJ and Joseph E Stiglitz (1980). “On the impossibility of informationally efficient markets”. In: *The American Economic Review* 70.3, pp. 393–408.

Grüne-Yanoff, Till (2009). “Learning from minimal economic models”. In: *Erkenntnis* 70.1, pp. 81–99.

Hands, D Wade (1997). “Caveat Emptor: Economics and Contemporary Philosophy of Science”. In: *Philosophy of Science*.

Hausman, Daniel M (1992). *The Inexact and Separate Science of Economics*. New York: Cambridge University Press.

— (2008). “Why Look Under the Hood?” In: *The Philosophy of Economics: An Anthology,Third Edition*. Ed. by Daniel M Hausman. Cambridge University Press.

Hausman, Daniel M and Michael S McPherson (2008). “The Philosophical Foundations of Mainstream Normative Economics”. In: *The Philosophy of Economics: An Anthology,Third Edition*. Ed. by Daniel M Hausman. Cambridge University Press.

Hicks, Daniel J (2016). “Bibliometrics for Social Validation”. In: *PLoS ONE* 11.12, e0168597–15.

Hong, Lu and Scott E Page (2004). “Groups of diverse problem solvers can outperform groups of high-ability problem solvers”. In: *Proceedings of the National Academy of Sciences of the United States of America*, pp. 16385–16389.

Katzav, Joel (2014). “The epistemology of climate models and some of its implications for climate science and the philosophy of science”. In: *Studies in History and Philosophy of Modern Physics* 46.PB, pp. 228–238.

Kitcher, Philip (1990). “The Division of Cognitive Labor”. In: *The Journal of Philosophy* 87.1, pp. 5–22.

Kuhn, Thomas S (1985). *The Copernican Revolution.* Planetary Astronomy in the Development of Western Thought. Harvard University Press.

Lenker, Lagretta (1999). “Why? Versus Why Not?: Potentialities of Aging in Shaw’s Back to Methuselah”. In: ed. by Sara M. Deats. Westport, CT: Praeger, pp. 47–59.

Lloyd, Elisabeth A (2009). “Varieties of Support and Confirmation of Climate Models”. In: *Aristotelian Society Supplementary Volume* 83.1, pp. 213–232.

Martini, Carlo and Manuela Fernández Pinto (2017). “Modeling the social organization of science”. In: *European Journal for Philosophy of Science* 7.2, pp. 221–238.

Massey, Douglas S and Nancy A Denton (1993). *American Apartheid.* Segregation and the Making of the Underclass. Harvard University Press.

Merton, Robert King (1957). “Priorities in Scientific Discovery: A Chapter in the Sociology of Science”. In: *American Sociological Review* 22.6, pp. 635–659.

Mirowski, Philip (2004). “Chapter 2: On Playing the Economics Cards in the Philosophy of Science: Why It Didn’t Work for Michael Polanyi”. In: *The Effortless Economy of Science?* Durham, N.C.: Duke University Press.

Morrison, Margaret (2015). *Reconstructing Reality: Models, Mathematics, and Simulations.* Oxford University Press.

Muldoon, Ryan and Michael Weisberg (2010). “Robustness and idealization in models of cognitive labor”. In: *Synthese* 183.2, pp. 161–174.

Orzack, Steven Hecht and Elliott Sober (1993). “A critical assessment of Levins’s the strategy of model building in population biology (1966)”. In: *The Quarterly Review of Biology* 68.4, pp. 533–546.

Perović, Slobodan et al. (2016). “Optimal research team composition: data envelopment analysis of Fermilab experiments”. In: *Scientometrics*, pp. 1–29.

Reijula, Heikki Samuli and Jaakko Erkinpoika Kuorikoski (2018). “Modeling epistemic communities”. In: *The Routledge Handbook of Social Epistemology*, pp. 1–18.

Romero, Felipe (2016). “Can the behavioral sciences self-correct? A social epistemic study”. In: *Studies in History and Philosophy of Science* 60.C, pp. 55–69.

Rosenstock, Sarita, Justin Bruner, and Cailin OConnor (2017). “In Epistemic Networks, Is Less Really More?” In: *Philosophy of Science*, pp. 234–252.

Schelling, Thomas C (1969). “Models of segregation”. In: *The American Economic Review* 59.2, pp. 488–493.

Simon, Herbert (2008). “Testability and Approximation”. In: *The Philosophy of Economics: An Anthology,Third Edition.* Ed. by Daniel M Hausman. Cambridge University Press, pp. 179–182.

Strevens, Michael (2003). “The role of the priority rule in science”. In: *The Journal of Philosophy*, pp. 55–79.

Sugden, Robert (2008). “Credible Worlds: The Status of Theoretical Models in Economics”. In: *The Philosophy of Economics: An Anthology,Third Edition.* Ed. by Daniel M Hausman. Cambridge University Press, pp. 476–509.

Thicke, Michael (2017). “Prediction Markets for Science: Is the Cure Worse than the Disease?” In: *Social Epistemology* 31.5, pp. 451–467.

Thoma, Johanna (2015a). “The Epistemic Division of Labor Revisited”. In: *Philosophy of Science* 82.3, pp. 454–472.

— (2015b). “The epistemic division of labor revisited”. In: *Philosophy of Science* 82.3, pp. 454–472.

Thompson, Abigail (2014). “Does Diversity Trump Ability?” In: *Notices of the American Mathematical Society* 61.09, pp. 1024–7.

Weisberg, Michael (2006). “Robustness Analysis”. In: *Philosophy of Science* 73.5, pp. 730–742.

— (2013). *Simulation and Similarity.* Oxford: Oxford University Press.

Weisberg, Michael and Ryan Muldoon (2009). “Epistemic Landscapes and the Division of Cognitive Labor”. In: *Philosophy of Science* 76.2, pp. 225–252.

Winsberg, Eric (2006). “Models of Success Versus the Success of Models: Reliability without Truth”. In: *Synthese* 152.1, pp. 1–19.

Zollman, Kevin JS (2010). “Social network structure and the achievement of consensus”. In: pp. 1–23.

Zollman, KJS (2007). “The communication structure of epistemic communities”. In: *Philosophy of Science* 74.5, pp. 574–587.