THE ROLE OF MODELING IN IMPROVING SCIENCE

AYDIN MOHSENI

Abstract. Recent developments in philosophy of science and metascientific research increasingly involve the use of formal models in analyzing the social structure and distinctive epistemic properties of science. This trend is underpinned by the implicit assumption that these models can be used to make meaningful inductive inferences into the nature of such structures and properties. We critically examine several substantive lines of this research. Our analysis focuses on identifying the insights and challenges posed by the research and indicating pressing open questions.

1. Introduction

Models can be used to understand and improve science. Something like this claim implicitly motivates much recent modeling work both in philosophy of science and in metascience.[1]

We proceed as follows. In §1, , we review several recent case studies of scientific modeling, and their discussion in the literature. In the second half of the chapter and explore the puzzles and principles these studies suggest for the practice of modeling science.

2. Lines of research

To model a system is to answer a series of questions regarding how to represent it. Which aspects of the target system should be included and which excluded? Of the features included, should they be formulated as quantities (categorical, ordinal, ratio, or interval), or as relations (synchronic or diachronic, deterministic or stochastic)? One way to organize the various lines of research that use models with the aim of improving science is by their broad answer to these questions, and hence the general class of the models they deploy in their analyses.

2.1. **Bandit model.**

2.2. **Filter models.**

2.3. **Epistemic landscape models.**

[1] For recent examples of models of science see Arvan et al. (forthcoming), Heesen (2023), Zollman (2018), Mohseni et al. (2023), Weatherall et al. (2020), **?**, Heesen and Romeijn (2019), Heesen (2018a), Grimes et al. (2017), Holman and Bruner (2017), and Smaldino and McElreath (2016).

2.4. **Evolutionary models.**

2.5. **Invisible hand models.**

3. General themes

3.1. **Independence thesis.**

3.2. **The importance of social structure.**

3.3. **The importance of non-epistemic incentives.**

4. Challenges

5. Open questions

5.1. **Filter models.** This The first case study is on work presented in Zollman (2007) and involves a model of the production of scientific findings which has been used to understand the impacts of methodological bias on the replication rates and reliability of published findings, and has figured prominently in debates regarding the merits of proposals for changing scientific practice.

The second case study is on work presented in Ioannidis (2005) and involves a model of the communication of scientific findings across social networks which has been used to investigate the impact of the structure of social networks on the incidence of premature lock in to false consensus; both theoretically and in the analyses of historical episodes.

In both cases models are used to explain some phenomenon of interest in science; and each points to some previously underappreciated epistemic dynamic. The explanatory burdens they take on, however, are instructively distinct. In the former case, a how-probably explanation is provided for the reliability of certain research findings; in the latter case, a how-possibly explanation is proffered for certain cases of premature lock-in to false consensus.

Much recent work in modeling science follows the formulas illustrated by our case studies. Typically, the phenomenon to be explained is motivated by some episode in the history of science. The model sets up conditions that reproduce something like the phenomenon of interest, such as low study reliability or premature lock in to consensus. Within the model, the causes of the phenomenon can then be identified, such as base rates and bias or relative rates of information sharing. From this, a potential explanation is proffered for the phenomena of interest.

From the causal relations identified within the model, we are invited to make some sort of inductive inference regarding causal relations in real-world scientific activity. The inference can be fairly modest, suggesting the causes identified

merely as candidate hypotheses, that is, as possibly at play in the real world;[2] which is still worthwhile in so far as they were not previously imagined as such. For how-probably explanations, the inference is characteristically stronger, suggesting the causes identified are 'likely' attributable in part or in whole. In Bayesian terms, one might think of the distinction between adding a hypothesis to the space of hypotheses of some boundedly rational agent versus proffering a compelling likelihood ratio for one or another hypothesis.

A key challenge of this sort of modeling is to understand what, if anything, should be learned from a model. We will use the case studies for rich examples in trying to address this question.

6. The filter model of science

Ioannidis (2005) popularized what might be called the filter model of science in his article, "Why most published research findings are false." In it, he anticipated the high replication failure rates that would be discovered in the replication crisis in the social and biomedical sciences.[3]

6.1. **The model.** Science might be conceived of as a process of filtration. (See Figure 1.) At first, researchers in a field or sub-field of science begin with the set of hypotheses they can formulate. These are the predictions of theory, past studies, conjectures, and common sense. These hypotheses are then put to test. Some pass these tests or acquire compelling evidence in their favor; others do not. Those that make it through this filter are submitted to the next: journals, conferences proceedings, textbooks, and the annals of scientific knowledge. These are the finding of a field at a time.

This process of filtration can be made mathematically precise. On doing this, one might ask clear questions about the relationship between various research practices, protocols, and paradigms and the properties of the scientific literature they produce.

Properties of the literature might include: the fraction of statistically significant hypotheses that turn out to be false, i.e., the false discovery rate; the fraction of non-significant hypotheses that turn out to be true, i.e., the false omission rate; the expected ratio of exaggeration of reported effect sizes to true effect sizes, i.e., the magnitude exaggerations ratio; or the expected impact study outcomes will have on important decisions, i.e., the decision-relevant informativeness of studies.

[2] For discussion of how-possibly, how-probably, and how-actually explanations, see Resnik (1991), and Šešelja (2023).

[3] For philosophical examinations of social and epistemic issues involved in the replication crisis see: Romero (2019, 2020); Romero and Sprenger (2020); Heesen (2018b); Bruner and Holman (2019); Bright (2017); Bird (2020); Devezer et al. (2019); Baumgaertner et al. (2019); and Machery (2020).

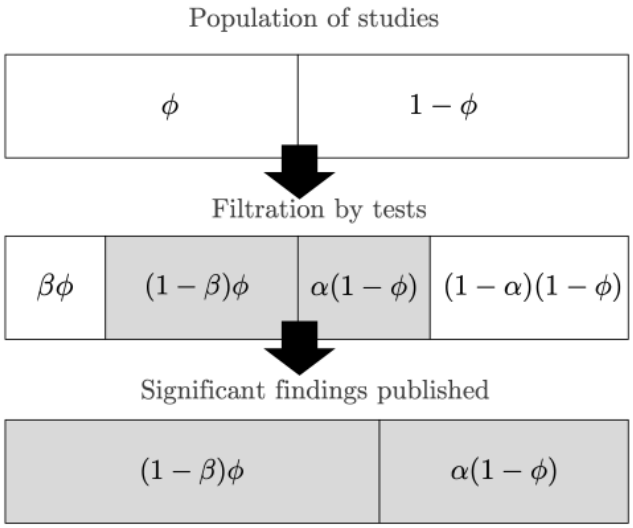


FIGURE 1. A visualization of an instance of the filter model of science where a population with fraction ϕ true initial hypotheses are submitted to tests with given type I and type II error rates, given by α and β, respectively, and where only statistically significant findings are ultimately published.

In this model, we can represent many proposals for remedial intervention. For example: the proposal to publish null results addresses the bottom filter (see Figure 1); proposals to limit researcher degrees of freedom via preregistration, lower the conventional threshold for statistical significance, and increase study power address the middle filter; and ambitious proposals to improve theory address the initial filter representing the set of hypotheses formulated by a field at a time.

6.2. **Initial inferences from the model.** Ioannidis (2005) used the filter model to explain and estimate the positive predictive value (PPV) of various types of studies as a function of their pre-study odds (R), the type I error (α) and statistical power ($1-\beta$) of their tests, along with their methodological bias (u).

As the PPV of a population studies corresponds to the expected the fraction of statistically significant findings that are true, it provides an estimate of the rate of successful replication of study findings for a given type of study. So, for example, for adequately powered randomized control trials with good pre-study odds (1:1), decent power ($1-\beta = .8$), and little bias ($u = .1$), the PPV is .85.[4] In other words, one should expect 85% of statistically significant results to be true and so successfully replicated under rigorous testing. (See Figure 2, row 1.) In contrast, for discovery-oriented exploratory research with massive testing (e.g.,

[4] We are assuming a conventional significance threshold of $\alpha = .05$ for these examples.

1 − β	R	u	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

FIGURE 2. Estimates of the positive predictive power (PPV) of various types research findings for various combinations of power $(1 - \beta)$, ratio of true to not-true relationships (R), and methodological bias (u). Reprinted from Ioannidis (2005).

certain GWAS studies) which exhibit low pre-study odds (1:1,000), low power $(1 = \beta = .2)$, and high bias $(u = .2)$, one should expect 0.15% of statistically significant results to be true and hence replicable.

In this way, Ioannidis illustrates how methodological bias and the error rates of studies, in tandem with different base rates of true hypotheses should be expected to affect the positive predictive value, and hence the replication rates, of published study findings.

6.3. **Subsequent applications & critical response.** There have been several subsequent applications of the Ioannidis (2005) filter model of science. Maniadis et al. (2014) uses the filter model to argue that the number of researchers investigating a topic should be expected to increase the false discovery rate of the new published research findings, and apply this insight to question to estimate the replicability of studies on anchoring effects. Crane (2018) and, earlier, Williams (2019) use adaptations of the filter model to argue that proposals to lower the convention threshold of statistical significance from .05 to .005 may produce unintended consequences, in particular, increasing the false discovery rate of new published research findings. These results serve as part of a back-and-forth on whether to change, keep, or eliminate the significance threshold.[5] Mohseni (2023) uses an adaptation of the filter model to argue that the research practice of HARKing, or hypothesizing after results are know, is misunderstood, and that a correct explanation of HARKing identifies its interaction with the interaction of researcher judgement and the base rates of the population of hypotheses tested.

[5] Cf. Benjamin et al. (2018), Ioannidis (2018), Machery (2019), and Lakens et al. (2018)

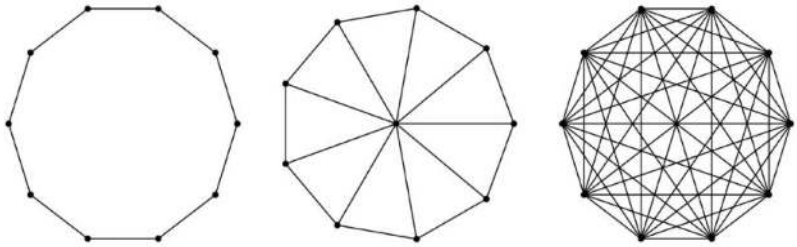


FIGURE 3. Three network configurations, illustrated here with six agents each. Reprinted from Zollman (2007).

7. The bandit model of science

Zollman (2007) first introduced the bandit model to philosophy of science. Using it, he explores the dynamics and reliability of epistemic communities–communities of learners, such as scientists–as they relate to the structures of communication of their social networks.

7.1. **The model.** Every action involves trade-offs; learning is no different. The bandit model[6] is perhaps the simplest interesting, formal expression of this fact. When combined with social networks, it can be used to explore certain epistemic and pragmatic trade-offs in the context of collective inquiry.

In its simplest form, the bandit model is composed of a decision problem with two possible actions: one better, one worse. However, the agent faces a problem: they do not yet know which action is better, and so must choose between exploring actions in order to better identify which is better (the 'explore' option) or focusing on the action they currently estimate to be better (the 'exploit' option). Hence, the game presents an explore-exploit tradeoff.

The basic idea can be adapted for modeling scientific inquiry. Consider a community of agents which might represent individual researchers or research labs–any unit that can choose to engage in running experiments to perform and thereby test of one several actions. The actions might be thought of as the choice of which medical treatment to administer, scientific theory to test, or research program to develop. The agents are related to one another by social relations, represented by edges on a network where with as agents, which indicates with whom each agent shares information (See Figure 3).

7.2. **Initial inferences from the model.** Zollman (2007) used the bandit model on networks to explain how less communication across a network of researchers might, counter-intuitively, produce more reliable consensus on the truth.

[6] The name come from the fact that slot machines, a mechanical instantiating of stochastic payoffs, are also known as "one-armed bandits".

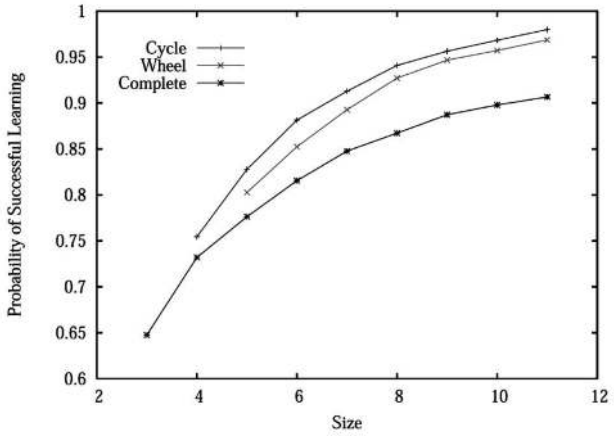


FIGURE 4. The Zollman effect: less connected networks can converge more reliably to the truth. Reprinted from Zollman (2007).

The historical episode used to motivate the model is that of the abandonment of the hypothesis of bacterial origins of peptic ulcer disease from the mid 1950s until the 1980s where further evidence revealed it to be true. In that case, a version of the historyRadomski et al. (2021) is that one research group disseminated finding that convinced the broader research community that bacterial origins were unlikely, and so gave up on that line of research.

In the bandit model of science, premature lock-in to false consensus occurs when the community as a whole is sufficiently confident that one action is better than the other that all members of the network cease investigating the ostensibly worse option.

A community can lower the rate of premature consensus by having some researchers continue to research the apparently less promising actions. One way this is achieved is to prevent research results from being universally shared so that different researchers will continue to have different assessments, at least for some time.[7]

This is the heart of the Zollman effect: less-connected networks of researcher are one way to preserve a diversity of opinions for longer, and so better hedge against premature lock-in to false consensus.[8]

7.3. **Subsequent applications & critical response.** A primary themes of subsequent research on the model has been the question of robustness. Rosenstock

[7] Of course, this comes at the cost of a trade-off with the rate at which the community converges to consensus.

[8] Later research has explored other ways that this diversity of opinions can be maintained Zollman (2012, 2013) , such as via the intransigence of some researchers (Zollman 2010; Wu and O?Connor 2023) (Holman and Bruner 2015), or variation in theoretical values between researchers (Zollman 2018).

et al. (2017) show that the Zollman effect obtains for a subset of the space of possible parameter values for the model, specifically where the success rates of the two actions are sufficiently similar, the population size is sufficiently small, and the amount of data collected in each round of play is sufficiently small. Borg et al. (2019) show that the Zollman effect does not obtain when certain substantial assumptions of the original bandit model are altered: when taking an action improves it future success rates, when researchers exhibit inertia in updating, when researchers can 'criticize' each other in a particular way, and when researchers consider both actions as equally good when their success rates are sufficiently similar but not identical. Radomski et al. (2021) question whether the Zollman effect actually explains the historical episode of peptic ulcer disease on the basis of textual analysis that suggest that the bacterial hypothesis had already been abandoned before the publication of the results that were supposed to have cause the premature abandonment of the hypothesis.

8. Discussion

With the preceding cases in had, we consider of the nature of inference from models of science, and propose best practices that might be adopted by future modelers of science.

8.1. **Inferences from models.** The analysis of a model can: suggest a novel hypothesis; demonstrate a possibility; clarify the implications of some line of reasoning; or, generally, shift our credences over the set of hypotheses under consideration.[9] Ultimately, the aim is to license some inference and influence the credal state of our reader regarding the nature of the scientific enterprise.

Getting clear on the nature of the model-world relation is hard. A compelling account of the nature of inference from idealized models is sketched in Robert Sugden's work on credible worlds (Sugden 2000, p. 1; 2009, p. 26. Sugden argues that models can be thought to describe counterfactual worlds, and that the the gap between the model worlds and the real world is bridged via inductive inference. That is, modeling results can support a sort of inductive inference over possible worlds. In so far as a result holds under a breadth of assumptions in idealized world described by our models, we might think it more likely that the result my hold of the actual world. In philosophy, Weisberg (2013) and Mayo-Wilson and Zollman (2021) present rich accounts of the role and epistemology of modeling.

Against the backdrop of this picture, one can ask, "What can be learned from this model?" Whatever the answer, we should get as clear as we can on this fact for ourselves and also for our readers. Whether the answer is "we learn some

[9] For the Bayesian angle–who already has all relevant hypotheses in her set of her hypotheses and has worked through their implications–it is perhaps all a shifting of credences.

phenomenon might arise in some way that was not previously appreciated" or "it is genuinely unclear what we learn, if anything", the ideal is that we understand and communicate as clearly as is in our power.[10]

To this end, I list a set of proposals for modeling epistemic communities in science. These proposals are certainly not original to me. Rather, they are the result of numerous conversations with other modelers of science over years. That said, I will inevitable have my own take on the recommendations.

Some of the proposed best practices, such as robustness analysis, are tacitly assumed and broadly practiced, if imperfectly. They are a part of the hidden curriculum of becoming a modeler. Others, such as explicitly stating the empirical assumptions and implications of our models are widely acknowledged as salutary, but more rarely practiced. All should be broadly known and discussed.

8.2. **The model/result distinction.** An important distinction is that between a result derived from a model versus the model itself. These can be conflated,[11] and are worth being clear on. The model itself is just the formal, computational, or even physical structure being analyzed. On its own, The model does not constitute a result; the analysis of a specific instantiation of the model.

For example, consider a simple, game-theoretic model: a normal-form game. Formally, the game is the mathematical object, $(N, (A_i), (u_i))$, consisting of a finite set, N, representing the players, a set, (A_i), representing the actions available to each player i, and the set, (u_i), of the utility functions $u_i : A \to \mathbb{R}$ for each player i where $A = \Pi_i \Delta A_i$ is the set of strategy profiles.

One might consider a specific instantiation of the model: a 2-player stag hunt. Now, we have $N = \{1, 2\}$, $(A_i) = \{A_1, A_2\}$ where $A_i = \langle S, H \rangle$, with payoffs satisfying $u_i(S, S) > u_i(S, H) > u_i(H, H) > u_i(S, H)$, for $i = 1, 2$.

With the game specified, we can derive a result. For example, we can find the Nash equilibria of the game–that is, the set of strategy profiles $A^* \in A$ which constitutes a fixed point under unilateral best response for ever player. Here, we find that the game exhibits two equilibria in pure strategies: (H, H) and (S, S).

Notice that the game and the result are distinct. Indeed, this same game can yield a variety of results when analyzed using various equilibrium concepts. One might consider the set of trembling-hand perfect equilibria (these will be the same for this game), evolutionary stable strategies (also the same), introduce pre-play signals and consider the set of correlated equilibria (of which there will continuum-many), explicitly introduce a deterministic dynamics, such as the replicator equations and consider the set of fixed points (the original Nash states will

[10] For skeptical discussions of the nature of modeling inference, see Šešelja (2019) and Thicke (2020).

[11] Indeed, I am guilty of this.

be asymptotically stable, but now there will be an unstable mixed state between them), or the Fermi process and consider the set of stochastically stable states (which, under the assumption of vanishing mutation, will consist in only one of the original Nash states).

Here, the *model* is the normal form game which describes the underlying interaction structures, and the *result* is that, when the model is instantiated in a particular way, and analyzed using some solution concept, some particular states are extracted and interpreted as predictions of the model.

These are statements about mathematical objects: model worlds and their properties. More work needs to be done to specify how the model system is meant to apply to some aspect of some other target system.[12] Notice that in distinguishing a model from model results, we can clarify distinct inferences: inference from the model to the target system, inference from the model result to the target system.

For example, consider toy model of the game. One might look at some real-world phenomena and make the following claim: the interaction is well-modeled as a simple strategic game, perhaps because it is a one-off interaction between approximately rational agents, but it is not a stag hunt with two Nash equilibria because the preferences of the agents which encode their choice behavior do not appear to satisfy the right inequalities, and so on.

Recall our cases from §6 and §7: the filter model and bandit model of science. The original articles using each model asked considered specific results derived from their those models.

In the first case, we were shown that, within a model of scientific production, under certain conditions,[13] we should expect the low replicability rates of published study findings. Here, we can ask distinct questions about the model in general (e.g., whether the behavior of the process of scientific production in general is well-captured by a linear filtration process as in Figure 1) and about a specific result (e.g., whether mean methodological bias of discovery-oriented exploratory research with massive testing close to the 0.8 number assumed for the 0.001 PPV result in Table Figure 2).

In the second case, we were shown that, within a model of social learning, under certain conditions[14] less-connected epistemic networks were more reliable. Here, we can ask of the model, as in Frey and Šešelja (2020), whether the assumption of fixed success rates in the model is apt. Or we could ask of the result that lower

[12] See Giere (2004) and Weisberg (2013) for accounts of the nature of the model-world relationship.

[13] When base rates of true alternative hypotheses are low, bias is high, and there is substantial publication bias.

[14] When the difference in success rates between actions was sufficiently small, the amount of data collected per round sufficiently small, and the population sufficiently small.

connectivity leading to higher reliability, as in Rosenstock et al. (2017), what the range is of parameter values of the model for which the result obtains.

These are all distinct failure modes to consider. A model and its results may not be apt for a given target system. A model may be generally apt for a target system but a result derived from that model may not, e.g., because it is non-robust within the model.[15] It is even possible a result may be apt for a target system whereas the model is generally not, e.g., because the model is apt for the system within only the limited regime of behavior in which the result happens to obtain.

8.3. **Robustness analysis.** In writing up our models, we should provide an analysis of reasonable expectations for each of what might be called the internal and external predictive validity of the model.[16]

Internal validity pertains to a result-model relation. Characteristically this involves demonstrating the *parametric robustness* and *structural stability* of a given result drawn from a model, where the parametric robustness of a result is the range of parameter values within the model under which the result obtains, and the dynamical robustness of the result is determined by whether the result obtains under perturbation of the underlying dynamics themselves.[17]

Consider a model composed of a system of differential equations, the Lyapunov stability of a state establishes a modicum of parametric stability by showing that arbitrary perturbations from the state will tend to return to the state. This can be extended by consider the size of the basin of attraction of the stable state, and extended again by considering epsilon perturbations of the model parameters, and again by a full sweep of the model parameter values. In contrast, the the notion of structural stability of a result, which asks whether its qualitative dynamics are preserved under perturbation of the dynamics itself, provides a measure of dynamical robustness. This can be extended by showing that a range of distinct dynamics with different underlying assumptions reproduce the same result.[18]

In the case of the bandit model of science, Rosenstock et al. (2017) and Frey and Šešelja (2020) are exploring the parametric and structural robustness of the (Zollman 2007) results, respectively. And in the case of the filter model of science,

[15] See Frey and Šešelja (2018) for a discussion of robustness in agent-based models of science.

[16] The terms proposed here are loosely analogous their counterparts in the design of empirical experiments. There, the internal validity of an experiment is a notion of how well the experiment is setup to measure and control for the desired empirical variables, and external validity of an experiment is a notion of how well the study findings can be generalized to contexts outside of the experimental setup.

[17] For excellent discussions of robustness under a range of dynamics see Skyrms (2000) and (Sandholm 2010, chs. 7, 8, and 12).

[18] For a result unifying several evolutionary dynamics, including the replicator equations and Lotka–Volterra equations, see Page and Nowak (2002). For the mean-field relationship between the replicator equations and reinforcement learning see Benaïm and Weibull (2003). For the relationship between the replicator equations and Bayesian inference see Harper (2010).

Machery (2020) questions the parametric stability of the replication rate backfire result in Crane (2018).

External validity applies to the model-world relation. This involves identifying the relation of the various aspects of the model and potential real-world target systems, and discussing the conditions under which the assumptions regarding these representational relationships should be expected to support or undermine the predictive validity of the model.

At present, the modeling work of philosophers of science rarely involves experimental testing to determine the external validity of our models.[19] By far the best critical analysis regarding the lack of empirical corroboration of formal models in philosophy of science thus far is to be found in Machery (2023).

8.4. **Proofs vs. simulations.** The analysis of models proceeds via proof or simulation. Proofs can often license characteristically stronger, deductive inferences about model behavior, but can be less informative or practically impossible. Simulations can provide information about systems where proof may be difficult, but typically license characteristically weaker, inductive inferences about model behavior.

Strength of inference about model behavior equates to neither to the strength of inference about a target system nor the import of one's results. That said, in doing interdisciplinary work, it can be helpful to know differences in disciplinary expectations regarding publishable work. It is important to know that in some disciplines, like microeconomics, deductive proofs may be expected for publication, while in other fields, like sociology, well-done simulation studies can suffice.

A common project workflow for modelers involves going back and forth between both methods of analysis. The following order of operations is illustrative: brainstorm and write out a model capturing some core idea; construct a computational version of the model in one's preferred language (e.g., NetLogo, Python, R, Wolfram Language); explore the behavior of the model using simulations; simulation results suggest that a pattern of interest probably holds over a range of conditions; attempt to prove analytic results showing the pattern definitely holds under precise conditions. And repeat.

8.5. **Choices of agents.** Models of science can contain an array of agents: researchers, reviewers, laboratories, regulatory bodies, corporations, and the public. When producing a model including such agents, we are confronted with the question of how best to mathematically represent them.

[19] Though there are recent exceptions: cf Bruner et al. (2018), Mohseni et al. (2021) and see Rubin et al. (2019) for an resource for experiment design particularly suited for modelers in philosophy.

The choice is not an easy one, and can be confusing to those unfamiliar with range of options. Bluntly put, there no perfect choices; there is no simple set of equations that accurately capture human behavior across a breadth of conditions. Complex statistical and machine learning models can achieve greater predictive accuracy, but they do so at the cost of the simplicity required for legible explanations. The types of models we will typically be using achieve some semblance of simplicity, but do so at the cost of quantitatively accurate predictions.

- *Ideal Bayesian agent.* An ideal Bayesian agent model is constituted as follows: she has probability function over an algebra of possibilities which encodes her beliefs about her learning situation, and she learns proposition by conditioning on them via Bayes' rule; she has a utility function which encodes her choice behavior and she acts by choosing a strategy which maximizes her expected utility; her beliefs are closed under deduction in the sense that they she knows the logical relations between the elements of her algebra. She assign utilities every outcome she might encounter and probabilities to every proposition she can conceive. Typically, she is also committed to epistemic principles such as the requirement of total evidence and possibly to some version of the principal principle.

  As a model of scientists, ideal Bayesian agent is unrealistically rational, computationally demanding, sometimes intractable, and so is more seldom the chosen model for agents. Instead, one chooses one her more bounded cousins.[20]

- *Bounded Bayesian agents*: The Bounded Bayesian agent is the most common formalism used for modeling scientists. The reason is that, however imperfect, she finds herself between the implausibly complex ideal Bayesian agent and other implausibly simple agents, such as naive imitation or reinforcement learners.

  There are three main ways to an agents may be bounded: by having her decision rule deviate from expected-utility maximization; by restricting her utility model; by have her learning rule deviate from Bayesian conditioning; or by restricting her representation of her learning situation.

  In Zollman (2007), agents are bounded primarily by the choice of a particularly simple utility function. Agents are rational in the sense that each chooses the action that maximizes her expected utility, but are *myopic* in so far as each has utilities only over the next round of play. This is equivalent to setting the agent's discount factor for future rounds of play to zero.[21]

[20] Cf. Simon (1957).

[21] Fully Bayesian solutions for non-zero discount factors are known and can be computed for simple bandit problems. These solutions are given by a Gittins index (Gittins 1979). However,

Two alternative decision rules are: noisy best response, which introduce some probability of error into actions; and logit choice model (R. D. 1959), which allows the modeling of a range of behaviors from classical best response to replicator dynamics to randomness by varying a single parameter.

The agents are all also rational in the sense that each learns via Bayes rule. However, since they only care about maximizing the expected value of their next action, this dramatically simplifies the representation of their beliefs, as strategies over indefinite time horizons need not be considered.

In other network models of science, where agents share signals or evidence with one another,[22] agents may not take signals at face value, and may entertain second-order considering over the information they are receiving. Note that essentially no models have fully sophisticated who entertain hypotheses over how other agents' signals may have been influenced by all of that agent's other potential interactions over the entire course of play.

The Bayesian brain hypothesis has produced an interesting paradigm. In this, we us the Bayesian apparatus as a model of optimality, and then find the optimal solution under various constraints on computational and representational resources. For some recents developments in this project see Icard (2018, 2019).

- *Imitation learning agents.* Imitation is perhaps the fundamental form of social learning, and a keystone of human cultural evolution (Boyd and Richerson 1988). Imitation learners are modeled as choosing strategies by their observed success in others. Imitation dynamics can capture the expected qualitative dynamics of large populations in certain cases, and so can be can be appropriate in modeling the transmission of scientific norms and behaviors.[23] Note that with many imitation dynamics, such as the replicator equations, it is not the agent that is being modeled, but rather a population composed of a distribution of strategies, behaviors, or traits. For canonical treatments of imitation dynamics, see Schuster and Sigmund (1983), Hofbauer and Sigmund (1998).
- *Reinforcement learning agents.* Some form of reinforcement learning is perhaps the oldest and most ubiquitous form of learning in the tree of

no such solutions are known, in generality, for bandit problems played by multiple agents sharing information via social networks.

[22] Cf. Mohseni and Williams (2021), Weatherall and O'Connor (2018).

[23] For example, in Smaldino and McElreath (2016), and Grimes et al. (2017).

life.[24] Its form is simple, and combines belief and action into one: a reinforcement learner begins with a probabilistic disposition to take one of finitely-many actions, and upon taking an action and observing its payoff, she reinforces her disposition to take that action in the future in proportion to the payoff received.[25] Note that reinforcement learning provides both a model for belief and for action–the agent has probabilistic dispositions to act, and these are modified through success and failure. While some degree of reinforcement learning is almost always at play in human learning, it is typically too primitive to serve as the model encompassing researcher belief and action.

These are a set of key dynamics, but there are more Sandholm (2010). The upshot remains: we tend to think that humans are in some sense 'between' Bayesian angels and primitive imitation and reinforcement learners. Whatever your choice of agent, and however it is bounded, it is worth thinking carefully about how your choices should affect the inferences you draw from your model.

8.6. **The virtue of adapting established models vs. developing wholly new ones.** A common mistake for new modelers is to reinvent the wheel. Often, the reinvention is inferior to extant models in the literature which have stood the test of time. This is not to say that one should never create entirely new models. This can be an important contribution. However, new modeling work benefit from understanding the modeling work that has come before.

There are other virtues in adapting existing models. Adapting an existing model will help other researchers familiar with the relevant literature to better understand one's results–its import, internal validity, and implications, generally. Additionally, adapting existing models that have been tested experimentally means that the experimental validity of one's model is better understood.

An all-too-common failure mode of modeling work is to deploy a bespoke model with no clear relationship to any existing literature, containing quantities that are meaningless,[26] with assumptions that are not fully understood or spelt out, and so to produce results whose implications are opaque.

8.7. **The interpretation of variables and meaningfulness.** Every modeler should have a working understanding of what the numbers used in their models mean. The study of the meaning of such numbers is measurement theory (Krantz et al. 1971). Are the measures involved invariant under positive affine transformations, such that they can be thought of as on an interval scale? Or merely

[24] Cf. Erev and Roth (1998).
[25] See (Huttegger 2017, ch. 2) for more on bounded rationality and learning.
[26] In the measurement-theoretic sense of meaningfulness discussed in §8.7.

under scalar transformations, such that they are on a ratio scale? Does the zero-point mean anything? Are there meaningful comparisons between the quantities within the model? What, if any, are the mappings of the quantities in a model to structures in the world?

There are practical implications of these questions. for example, in game theory, if one is interpreting the payoffs of a game in terms of the utilities of a rational agent (which, in decision frameworks such as that of Savage (1954) are invariant under positive affine transformation) it will yield different results and allow different transformations[27] than if the payoffs are to be interpreted in terms of the fitness of some type in the evolutionary context[28] (which, in finite population models such as the Moran process cannot be negative and may not even be invariant under scalar transformations).

8.8. **Making assumptions & implications clear.** We typically require that scholars situate their model within the literature. This helps the reader to understand the state of knowledge on the relevant question and to situate the paper in relation to that state of knowledge. It also helps the subject-area expert in to assess the background knowledge of the author(s) regarding the state of knowledge and their own contribution.

Similarly, we should require that, as modelers, we lay out our modeling assumptions and vet them against extant empirical literature. This will help the reader to understand if and when key assumptions are plausibly satisfied; and when they are not.

We should also make the implications of our models clear. In particular, we should indicate what further evidence would increase or decrease our credences in the applicability of our modeling results.

Just as we have *methods* sections in experimental papers, a reasonable proposal is that modeling papers in philosophy of science include 'assumptions' and 'implications' sections as standard elements. The inclusion of these sections as standard in our papers can remind us to think clearly and deeply about these topics, share what we are thinking with the reader, facilitate understanding of the inferential warrant of our results, and enable criticism of assumptions and inferences.

References

Arvan, M., L. K. Bright, and R. Heesen (forthcoming). Jury theorems for peer review. *The British Journal for the Philosophy of Science*.

[27] For a review of decision frameworks and their properties, see Fishburn (1981).
[28] Cf. Rubin et al. (2019) for analysis of the implications of various fitness representations.

Baumgaertner, B., B. Devezer, E. O. Buzbas, and L. G. Nardin (2019). Openness and reproducibility: Insights from a model-centric approach. *PLos ONE 14*(5), e0216125.

Benaïm, M. and J. Weibull (2003). Deterministic Approximation of Stochastic Evolution in Games. *Econometrica 71*(3), 873–903.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E. J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson (2018). Redefine statistical significance. *Nature Human Behaviour 2*, 6–10.

Bird, A. (2020). Understanding the Replication Crisis as a Base Rate Fallacy. *The British Journal for the Philosophy of Science 0*, 1–31.

Boyd, R. and P. Richerson (1988). *Culture and the Evolutionary Process*. Biology, Anthropology, Sociology. University of Chicago Press.

Bright, L. K. (2017). On fraud. *Philosophical Studies 174*, 291–310.

Bruner, J., C. O'Connor, H. Rubin, and S. M. Huttegger (2018). David lewis in the lab: experimental results on the emergence of meaning. *Synthese 195*(2), 603–621.

Bruner, J. P. and B. Holman (2019). Self-correction in science: Meta-analysis, bias and social structure. *Studies in history and philosophy of science 78*, 93 – 97.

Crane, H. (2018). The impact of p-hacking on "redefine statistical significance". *Basic and Applied Social Psychology*.

Devezer, B., L. G. Nardin, B. Baumgaertner, and E. O. Buzbas (2019, 05). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE 14*, 1–23.

Erev, I. and A. E. Roth (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review 88*(4), 848–881.

Fishburn, P. C. (1981). Subjective expected utility: A review of normative theories. *Theory and Decision 13*(2), 139–199.

Frey, D. and D. Šešelja (2020). Robustness and idealizations in agent-based models of scientific interaction. *The British Journal for the Philosophy of Science 71*(4), 1411–1437.

Frey, D. and D. Šešelja (2018). What is the epistemic function of highly idealized agent-based models of scientific inquiry? *Philosophy of the Social Sciences 48*(4), 407–433.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science 71*(5), 742–752.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological) 41*(2), 148–164.

Grimes, D. R., C. T. Bauch, and J. P. Ioannidis (2017). Modeling science trustworthiness under publish or perish pressure. *bioRxiv*.

Harper, M. (2010). The replicator equation as an inference dynamic.

Heesen, R. (2018a). Why the reward structure of science makes reproducibility problems inevitable. *Journal of Philosophy 115*(12), 661–674.

Heesen, R. (2018b). Why the reward structure of science makes reproducibility problems inevitable. *Journal of Philosophy 115*(12), 661–674.

Heesen, R. (2023). Cumulative advantage and the incentive to commit fraud in science. *The British Journal for the Philosophy of Science*.

Heesen, R. and J. Romeijn (2019). Epistemic diversity and editor decisions: A statistical matthew effect. *Philosophers' Imprint 19*.

Hofbauer, J. and K. Sigmund (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Holman, B. and J. Bruner (2017). Experimentation by industrial selection. *Philosophy of Science 84*(5), 1008–1019.

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

Huttegger, S. M. (2017). *The Probabilistic Foundations of Rational Learning*. Cambridge University Press.

Icard, T. (2019, 11). Why Be Random? *Mind 130*(517), 111–139.

Icard, T. F. (2018). Bayes, bounds, and rational analysis. *Philosophy of Science 85*(1), 79–101.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine 2*(8), e124.

Ioannidis, J. P. (2018). The proposal to lower P value thresholds to .005. *Journal of the American Medical Association 319*(14), 1429–1430.

Krantz, D. H., P. Suppes, and R. D. Luce (1971). *Foundations of measurement*, Volume 2. academic press.

Lakens, D., F. G. Adolfi, C. J. Albers, F. Anvari, M. A. J. Apps, S. E. Argamon, T. Baguley, R. B. Becker, S. D. Benning, D. E. Bradford, E. M. Buchanan, A. R. Caldwell, B. Van Calster, R. Carlsson, S.-C. Chen, B. Chung, L. J. Colling, G. S. Collins, Z. Crook, E. S. Cross, S. Daniels, H. Danielsson, L. DeBruine, D. J. Dunleavy, B. D. Earp, M. I. Feist, J. D. Ferrell, J. G. Field, N. W. Fox, A. Friesen, C. Gomes, M. Gonzalez-Marquez, J. A. Grange, A. P. Grieve, R. Guggenberger, J. Grist, A.-L. van Harmelen, F. Hasselman, K. D. Hochard, M. R. Hoffarth, N. P. Holmes, M. Ingre, P. M. Isager, H. K. Isotalus, C. Johansson, K. Juszczyk, D. A. Kenny, A. A. Khalil, B. Konat, J. Lao, E. G. Larsen, G. M. A. Lodder, J. Lukavský, C. R. Madan, D. Manheim, S. R. Martin, A. E. Martin, D. G. Mayo, R. J. McCarthy, K. McConway, C. McFarland, A. Q. X. Nio, G. Nilsonne, C. L. de Oliveira, J.-J. O. de Xivry, S. Parsons, G. Pfuhl, K. A. Quinn, J. J. Sakon, S. A. Saribay, I. K. Schneider, M. Selvaraju, Z. Sjoerds, S. G. Smith, T. Smits, J. R. Spies, V. Sreekumar, C. N. Steltenpohl, N. Stenhouse, W.Swiktkowski, M. A. Vadillo, M. A. L. M. Van Assen, M. N. Williams, S. E. Williams, D. R. Williams, T. Yarkoni, I. Ziano, and R. A. Zwaan (2018). Justify your alpha. *Nature Human Behaviour 2*(3), 168–171.

Machery, E. (2019). The alpha war. *Review of Philosophy and Psychology 12*(1), 75–99.

Machery, E. (2020). What is a replication? *Philosophy of Science 87*(4), 545–567.

Machery, E. (2023). Formal modeling in philosophy of science – let's be realistic!

Maniadis, Z., F. Tufano, and J. A. List (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review 104*(1), 277–90.

Mayo-Wilson, C. and K. J. S. Zollman (2021). The computational philosophy: simulation as a core philosophical method. *Synthese 199*(1), 3647–3673.

Mohseni, A. (2023). Harking: from misdiagnosis to misprescription.

Mohseni, A., C. O'Connor, and H. Rubin (2021). On the emergence of minority disadvantage: testing the cultural red king hypothesis. *Synthese 198*(6), 5599–5621.

Mohseni, A., C. O'Connor, and J. O. Weatherall (2023). The best paper you'll read today: Media biases and the public understanding of science. *Philosophical Topics*.

Mohseni, A. and C. R. Williams (2021). Truth and conformity on networks. *Erkenntnis 86*(6), 1509–1530.

Page, K. M. and M. A. Nowak (2002). Unifying evolutionary dynamics. *Journal of Theoretical Biology 219*(1), 93–98.

R. D., L. (1959). *Individual choice behavior. John Wiley.* John Wiley.

Radomski, B. M., D. Šešelja, and K. Naumann (2021). Rethinking the history of peptic ulcer disease and its relevance for network epistemology. *History and Philosophy of the Life Sciences 43*(4), 113.

Resnik, D. B. (1991). How-possibly explanations in biology. *Acta Biotheoretica 39*(2), 141–149.

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass 14*, e12633.

Romero, F. (2020). The Division of Replication Labor. *Philosophy of Science 87*(5), 1014–1025.

Romero, F. and J. Sprenger (2020). Scientific self-correction: the Bayesian way. *Synthese*.

Rosenstock, S., J. Bruner, and C. O'Connor (2017). In epistemic networks, is less really more? *Philosophy of Science 84*, 234–252.

Rubin, H., C. O'Connor, and J. Bruner (2019, March). Experimental economics for philosophers. In E. Fischer and M. Curtis (Eds.), *Methodological Advances in Experimental Philosophy* (1 ed.). Bloomsbury Academic.

Sandholm, W. H. (2010). *Population Games and Evolutionary Dynamics*. The MIT Press.

Savage, L. (1954). *The Foundations of Statistics*. Wiley Publications in Statistics.

Schuster, P. and K. Sigmund (1983). Replicator dynamics. *Journal of Theoretical Biology 100*(3), 533–538.

Simon, H. (1957). *Models of Man: Social and Rational; Mathematical Essays on Rational Human Behavior in Society Setting*. Continuity in administrative science. Wiley.

Skyrms, B. (2000). Stability and explanatory significance of some simple evolutionary models. *Philosophy of Science 67*(1), 94–113.

Smaldino, P. E. and R. McElreath (2016). The natural selection of bad science. *Royal Society Open Science 3*(9), 160384.

Sugden, R. (2000). Credible worlds: the status of theoretical models in economics. *Journal of Economic Methodology 7*(1), 1–31.

Sugden, R. (2009). Credible worlds, capacities and mechanisms. *Erkenntnis 70*(1), 3–27.

Thicke, M. (2020). Evaluating formal models of science. *Journal for General Philosophy of Science / Zeitschrift für Allgemeine Wissenschaftstheorie 51*(2), 315–335.

Šešelja, D. (2019). Some lessons from simulations of scientific disagreements. *Synthese 198*(Suppl 25), 6143–6158.

Šešelja, D. (2023). What kind of explanations do we get from agent-based models of scientific inquiry? In H. Andersen, T. Marvan, H. Chang, B. Löwe, and I. Pezlar (Eds.), *Proceedings of the 16th International Congress of Logic, Methodology and Philosophy of Science and Technology*.

Weatherall, J. O. and C. O'Connor (2018). Conformity in scientific networks. *Synthese*, 1–22.

Weatherall, J. O., C. O'Connor, and J. P. Bruner (2020). How to beat science and influence people: Policymakers and propaganda in epistemic networks. *The British Journal for the Philosophy of Science 71*(4), 1157–1186.

Weisberg, M. (2013, 01). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

Williams, C. R. (2019). How redefining statistical significance can worsen the replication crisis. *Economics Letters 181*, 65–69.

Wu, J. and C. O?Connor (2023). How should we promote transient diversity in science? *Synthese 201*(2), 1–24.

Zollman, K. J. S. (2007). The Communication Structure of Epistemic Communities. *Philosophy of Science 74*(5), 574–587.

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

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

Zollman, K. J. S. (2013). Network epistemology: Communication in epistemic communities. *Philosophy Compass 8*(1), 15–27.

Zollman, K. J. S. (2018). The credit economy and the economic rationality of science. *The Journal of Philosophy 115*, 5–33.