Forthcoming in: Hanne Andersen, Tomáš Marvan, Hasok Chang, Benedikt Löwe, and Ivo Pezlar (eds.), *Logic, Methodology and Philosophy of Science and Technology. Proceedings of the Sixteenth International Congress, Bridging Across Academic Cultures*. Rickmansworthe: College Publications, 2022.

What Kind of Explanations Do We Get from Agent-Based Models of Scientific Inquiry?

Dunja Šešelja

Eindhoven University of Technology, d.seselja@tue.nl

Abstract

Agent-based modelling has become a well-established method in social epistemology and philosophy of science but the question of what kind of explanations these models provide remains largely open. This paper is dedicated to this issue. It starts by distinguishing between real-world phenomena, real-world possibilities, and logical possibilities as different kinds of targets which agent-based models (ABMs) can represent. I argue that models representing the former two kinds provide how-actually explanations or causal how-possibly explanations. In contrast, models that represent logical possibilities provide epistemically opaque how-possibly explanations (Šešelja et al., 2022). While highly idealised ABMs in the form in which they are initially proposed typically fall into the last category, the epistemic opaqueness of explanations they provide can be reduced by validation procedures. To this purpose, an examination of results of simulations in terms of classes of models can be particularly helpful. I illustrate this point by discussing a class of ABMs of scientific interaction and the claim that a high degree of interaction can impede scientific inquiry.

Keywords: agent-based models, highly idealised models, epistemically opaque how-possibly explanation, robustness analysis, scientific interaction.

1 Introduction

Computer simulations in the form of agent-based models (ABMs) have become a well-established formal method in social epistemology and philosophy of science. Following a long tradition in biomedical and social sciences, this computational method had quickly proven itself useful in the study of social aspects of scientific inquiry in subjects ranging from the impact of different social networks on the efficiency of

knowledge acquisition and the division of cognitive labour all the way to research of the efficiency of scientific collaboration and studies of the norms that guide scientists facing disagreements. The primary advantage of using ABMs to examine such issues is that they allow us to study, in a controlled environment, how the various properties of individual agents representing scientists – such as their reasoning, decision-making, actions, and relations – bring about various phenomena on the level of the scientific community, such as the success or a failure of the community to acquire knowledge.

Despite their popularity, studies based on computer simulations often meet with sceptical reactions of researchers who use other approaches to the philosophy of science, such as for instance historical case studies. Their primary concern is that the proposed models are highly idealised, which raises the question of validity of any findings such models may deliver. In particular, the simplicity with which the ABMs tend to represent scientific inquiry commonly leads to doubts regarding their explanatory value, such as: 'Do these models explain anything, and if so, what exactly?' 'Surely, they cannot be taken as explanatory of complex scientific episodes, which include a myriad of epistemic and non-epistemic causal factors?'

In this paper, I want to address these concerns and explain the nature of explanations which ABMs provide. I start (in Section 2) by distinguishing three focal points in the research on ABMs of science: the development of highly idealised models, studies of their robustness, and discussions of the epistemology of agent-based modelling. This will allow me to situate the current contribution within the third of the above mentioned points. To examine the explanatory properties of highly idealised ABMs, I distinguish the different possible targets which ABMs can adequately represent, and then proceed to relate this classification to the types of explanation that can be inferred from each class (Section 3). I argue that highly idealised models that have not been validated provide *epistemically opaque how-possibly explanations*, that is, claims that express possible causal relationships although the conditions under which such relationships should hold are unclear. Further, I suggest that by the means of different validation procedures, ABMs can move from providing epistemically opaque explanations to *causal how-possibly explanations* (Section 4). I illustrate this point with a class of ABMs of scientific interaction and with a claim inferred on their basis, namely that a high degree of information flow can be detrimental to the efficiency of a scientific inquiry (Section 5). Section 6 then concludes the paper.

2 Research on ABMs of science

We can roughly distinguish three main directions in the research on ABMs of scientific inquiry developed within the philosophy of science. To explain the main questions raised within each of these focal points, let us first look at how the philosophical study of ABMs developed from other scientific domains.

Simulations of scientific inquiry are rooted in several parallel lines of research.[1] On the one hand, formal modelling was introduced into the philosophical study of social processes underlying scientific inquiry with the aim of gaining more precise insight into the tensions pervading scientific research, such as the tension between individual and group rationality or between epistemic and non-epistemic values.[2] That resulted in a number of analytical models, such as the model proposed by Goldman and Shaked (1991), which examined the relationship between the goal of one's professional success and promotion of truth acquisition, or Kitcher's models (1990, 1993), which tackled the division of cognitive labour against the background of individual rationality. These were later followed by several other proposals (e.g., Strevens, 2003; Zamora Bonilla, 1999; Zamora Bonilla, 2002).

Around the same time, computational methods entered the philosophical study of rational deliberation and cooperation in the context of game theory (Skyrms, 1990, 1996; Grim et al., 1998) and the study of opinion dynamics in social epistemology (Hegselmann and Krause, 2002, 2005). The latter, which focused on investigating how beliefs and opinions change within a group of agents, was studied by means of ABMs.

In a parallel development, agent-based modelling entered also the social sciences. In sociology of science, ABMs offered a novel way of analysing and explaining causal mechanisms underlying scientific inquiry, an approach that complemented the more entrenched method of quantitative empirical studies. The pioneering work of Gilbert (1997), aimed at simulating the structure of academic science, was closely related to a quantitative analysis of citation networks. Using a small number of simple assumptions, Gilbert's ABM was designed to reproduce certain quantitative relationships previously identified in empirical research (such as Lotka's Law concerning the distribution of citations among authors).

[1] For a recent overview of formal models of scientific inquiry and their role in philosophical literature, see Šešelja et al. (2020); for an overview of ABMs of scientific interaction see Šešelja (2022); for an overview of computational methods employed in philosophy, see Grim and Singer (2020) and Mayo-Wilson and Zollman (2021). For an earlier overview of ABMs of science, including both work done in sociology and in philosophy of science, see Payette (2012); for an overview of agent-based modelling and its role in social sciences and philosophy, see Klein et al. (2018); for a discussion of the use of computer models in science in general, see Imbert (2017).

[2] For an overview of economic approaches to social epistemology of science, which inspired discussions on the tension between the individual and group rationality, see Mäki (2005).

In contrast to ABMs developed in the sociology of science, which tended towards an integration of simulations and empirical studies used for their validation (cf. Gilbert and Troitzsch, 2005), a parallel trend of abstract and highly idealised ABMs emerged in other social sciences, such as economics and archaeology. Most prominently, Schelling–Sakoda models of social segregation (Sakoda, 1971; Schelling, 1971, 1978; see also Hegselmann, 2017) and Axelrod's models of cooperation (e.g., Axelrod, 1984, 1997; Axelrod and Hamilton, 1981) paved the ground for agent-based modelling in the study of various social phenomena. These two trends gave rise to two distinct methodological approaches to ABMs that came to be known as KIDS (*Keep it Descriptive, Stupid*) and KISS (*Keep it Simple, Stupid*) strategies. The KIDS approach aims at developing models which are descriptively adequate with respect to central features of the target phenomenon and at integrating ABMs and empirical studies. The KISS approach, on the other hand, aims at the development of simple, highly idealised models which are based on a minimal set of assumptions about agents and their environment but sufficient to capture certain regularities on the community level.[3]

The development of ABMs in the philosophy of science has largely followed the KISS approach. The influential works of Hegselmann and Krause (2006), Zollman (2007, 2010), Muldoon and Weisberg (2011), Weisberg and Muldoon (2009), Grim (2009), Grim et al. (2013), and Douven (2010), among others, kickstarted research into abstract ABMs of scientific inquiry. This marks the first focal point in the research on ABMs in the philosophy of science. The development of ABMs aimed at demonstrating the contribution of agent-based modelling to the study of questions posed by philosophers of science and social epistemologists, such as the impact of social networks or division of cognitive labour on the efficiency of inquiry. The emphasis was on exploratory insights rather than validity of the models or a detailed analysis of their explanatory features. For modellers endorsing the KISS approach, this aim continued to be central.

Others, however, recognised the limitations of this approach. On the one hand, it is generally acknowledged that highly idealised models are sufficient to provide a 'proof of the concept', for instance, to show that a certain causal relationship is in principle possible (Šešelja, 2021). Similarly, highly idealised models are capable of producing conjectures about causal mechanisms underlying real-world phenomena. On the other hand, abstract models typically lack validation procedures, such as robustness analysis or studies of their representational adequacy (Aydinonat et al., 2020). This makes it difficult to

[3] For the KISS strategy see, e.g., Epstein and Axtell (1996), Axelrod (1997), Hegselmann and Krause (2002), Epstein (2006); for the KIDS one, see Edmonds and Moss (2004).

assess whether and to what extent findings from these models can be considered informative of real-world scientific inquiries.

Such concerns gave rise to the second focal point in the research on ABMs in the philosophy of science: the study of robustness of previously developed models. To this end, previous models were adjusted and enhanced, resulting in what Aydinonat et al. (2020) called 'second generation models'.[4] The robustness analysis includes an examination of results delivered by a model with respect to changes in parameter values (sensitivity analysis) and changes to the idealising assumptions of the model (derivational robustness analysis). For example, with respect to Zollman's models (2007, 2010), Rosenstock et al. (2017) showed that the previously obtained results hold only for a small part of the relevant parameter space, while Frey and Šešelja (2020) and Borg et al. (2019) showed that Zollman's results do not obtain when some of the idealising assumptions are changed. Similar studies were conducted for Weisberg and Muldoon's (2009) model: others identified an error in the code of the model and critically assessed the robustness of results under different modelling assumptions (Alexander et al., 2015; Thoma, 2015; Pöyhönen, 2017; Pinto and Pinto, 2018).

Besides studies of robustness, enhancements of previously proposed ABMs have also led to their application to new research questions. For instance, a number of ABMs studying scientific polarisation, biases, or the spread of deceptive information were built on Zollman's work (see works by Holman and Bruner, 2015, 2017; O'Connor and Weatherall, 2018, 2019; Weatherall et al., 2018). Similarly, Weisberg and Muldoon's epistemic landscape model served as a starting point for various further studies: for instance, Balietti et al. (2015) studied the relationship between disciplinary fragmentation and scientific progress, Currie and Avin (2018) examined different types of scientific methods, while Harnagel (2018) and Avin (2019) focused on the mechanisms of allocation of research funding.

Finally, the third focal point in research on ABMs of science concerns the epistemology of agent-based modelling. What can we learn from ABMs? What kind of epistemic functions do they have? What are their limitations and prospects for future improvement? These interrelated questions have been examined in a number of studies. On the one hand, some have argued that unless ABMs are empirically embedded and validated, we will have a hard time ensuring their empirical adequacy (e.g., Martini and Pinto, 2016; Thicke, 2020; Bedessem, 2019; Frey and Šešelja; 2018, Šešelja, 2021; Politi, 2021). For instance, by using empirical data as the input for ABMs we can calibrate parameters in the model (for example,

[4] Research on ABMs in empirical sciences has followed a similar course. For instance, Thiele et al. (2014) identify two phases in their development: The first focused on gaining generic insights via ABMs rather than on their in-depth analysis. In the second phase, previously developed models are subjected to various types of robustness analyses with the goal of 'better mechanistic understanding of the model and on relating the model to real-world phenomena and mechanisms'.

Harnagel, 2018, used bibliometric data to this purpose). On the other hand, Mayo-Wilson and Zollman (2021) have argued that for some modelling purposes, such as illustrating that certain events or situations are possible, validation need not be necessary. Models can instead be justified by 'plausibility arguments' and by recourse to stylised historical case studies.

Central to the above discussion is the question of the epistemic purpose of a model. Aydinonat et al. (2020) have argued that this may be difficult to assess when examining a model in isolation. According to them, we instead ought to take a 'family-of-models perspective' and determine the contribution of an ABM using subsequent models that enable a better understanding of results delivered by the previous ones.[5] More precisely, Aydinonat and colleagues argue that we should view the ABMs as argumentative devices whose purpose is determined by the argumentative context in which they are used. An argument supported by a particular model can be further strengthened by analyses based on subsequent models.

This paper belongs to the third focal point in research on ABMs of science. While previous discussions examined the conditions under which ABMs can be explanatory of real-world phenomena, the question what kind of explanations highly idealised ABMs of science provide remained open. Using the perspective of Aydinonat et al. (2020), we can say that this boils down to the following questions: Can we use highly idealised ABMs of science to construct explanatory arguments, and if so, of what kind? An attempt to answer this question is the subject of the following section.

3 ABMs and their explanatory power

What can we learn from highly idealised ABMs of science and what exactly do they represent? The answer is far from trivial and it is closely related to the ongoing philosophical debate about the epistemic function of highly idealised or 'toy' models in empirical sciences (e.g. Alexandrova, 2008; Fumagalli, 2016; Grüne-Yanoff, 2009; Hoyningen-Huene, 2020; Nguyen, 2019; Reiss, 2012; Reutlinger et al., 2018, cf. also references in Footnote 8). While my aim is to address this question by focusing on ABMs in the philosophy of science, the bulk of this section is sufficiently general to apply to toy models in other disciplines as well. I start by distinguishing between different possible targets which models can adequately represent and I relate them to the different types of explanations a particular representation licenses. Then I turn to validation strategies, which help us move a model from one explanatory category

[5] Similar methodological approaches have been endorsed in the context of ABMs in the social sciences, see e.g. Page (2018), Kuhlmann (2021).

to another. Finally, I go back to the ABMs of science and examine how they are to be classified both before and after passing a certain validation procedure.

3.1 What do ABMs represent?

According to Bolinska (2013), 'A vehicle is an *epistemic representation* of a given target system if and only if it is a tool for gaining information about this system'. Here, information denotes those considerations which are not readily accessible by directly observing the target but can be understood via a particular vehicle, in this case by the means of a particular model. In the remainder of this article, whenever I speak of a model representing a target, I refer to an epistemic representation of the target. After distinguishing different types of targets, I will specify the types of explanations that their representation warrants.

One way of categorising the representational properties of ABMs is according to whether they represent actual or possible phenomena. On the one end of the spectrum, there are ABMs that represent real-world phenomena (Fig. 1).[6] These models were developed most prominently in urban planning and epidemiology, where they have been used for policy guidance. For instance, the UrbanSim (Waddell, 2002) set of models of urban planning was developed to guide urban policy and transportation investments. While UrbanSim was built based on empirical data, it was designed as a virtual experimental lab where various counterfactual scenarios can be represented and analysed (Bruch and Atwell, 2015). In other words, these models were built not merely to represent actual empirical processes, but also – and crucially – to model 'real-world possibilities', that is, scenarios that could take place once some factors are altered. This is the second kind of targets ABMs can represent. The capacity of models to represent possibilities is essential for drawing normative and descriptive conclusions from them, because it allows us to draw inferences about counterfactual dependencies concerning the purported target. For instance, models of herd immunity and disease spread, which are used to examine different policies of epidemics management, enable the acquisition of precisely this sort of knowledge (Epstein, 2009).

[6] This classification should not be taken as exhaustive since some issues are either lacking or require further disambiguation. For example, non-existent targets, which can be part of hypothetical modelling, may be physically impossible and yet informative of real-world phenomena and their possibilities (Weisberg, 2013, 121–122).

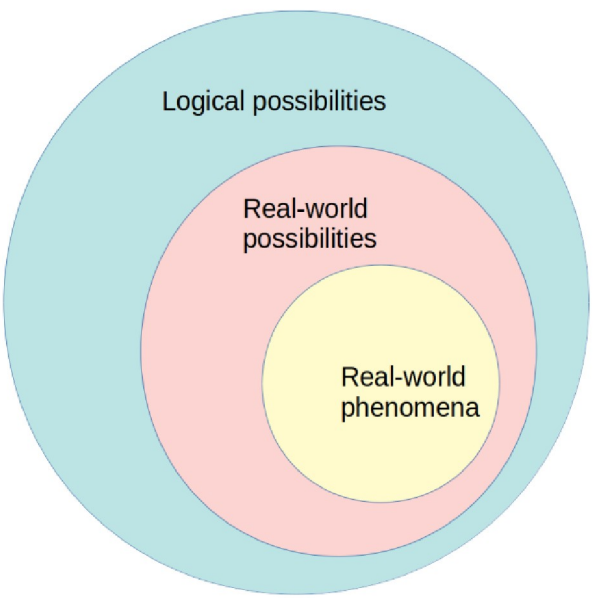

Figure 1: A simplified picture of different target phenomena represented by ABMs.

Models mentioned in the previous paragraph increase our explanatory understanding of the phenomena they represent in the sense of expanding our ability to make reliable 'what-if' inferences about them (Ylikoski, 2014). In contrast to such models, other simulations represent only logical possibilities. These are scenarios that may but need not correspond to any interesting real-world possibilities. Importantly, determination of whether they correspond to real-world possibilities – and if so, which ones – is an open question. Hence, these are models from which we cannot draw reliable 'what-if' inferences. As I argue below, highly idealised models upon their initial development typically belong to this category.

To make this classification more precise, let us look into the kind of explanations that each class warrants.[7]

3.2 How-possibly and how-actually explanations

In view of the above classification of the modelled targets, it is helpful to make a related distinction between *how-actually* explanations (HAEs), and *how-possibly* explanations (HPEs). While the former notion concerns explanations simpliciter, that is, accounts of how phenomena actually occur, the latter was introduced to cover accounts of possible ways in which phenomena can occur.[8] Following Verreault-

[7] Explanation of phenomena is certainly not the only epistemic function of ABMs. For other epistemic functions of ABMs see e.g. Edmonds et al. (2018), Epstein (2006), Frey and Šešelja (2018).

[8] The notion of HPE was introduced by Dray (1957) in the context of explanations in history. Subsequently, it became the subject of extensive debates in the literature on scientific modelling, especially in biology and social sciences (see, e.g., Bokulich, 2014, 2017; Forber, 2010, 2012; Hempel, 1965; Reydon, 2012; Ylikoski and Aydinonat, 2014). One can find different versions of this

Julien (2018), we can characterise HAEs as expressing propositions of the form '*p* because *q* (and initial conditions *c*)'. In contrast, HPEs express propositions of the form 'it is possible that: *p* because *q* (and initial conditions *c*)'. HPEs can express various types of modalities, such as mathematical or causal ones.

Based on the above, we can characterise the explanatory properties of ABMs as follows:

ABMs representing *real-world phenomena* and *real-world possibilities* provide one of the two following types of explanations:

- **HAEs** which express propositions of the form: '*p* because *q* and initial conditions *c*', where we know which conditions these are and we know that they hold for a particular empirical target.

- **causal HPEs**,[9] which express propositions of the form: 'It is *causally* possible that: *p* because *q* and initial conditions *c*.', where we know which conditions these are al, though we may not know whether they hold for a particular empirical target, or we know that they do not hold for that particular target.[10]

ABMs representing *logical possibilities* provide:

- ***epistemically opaque HPEs*** (ep-op HPEs),[11] which express propositions of the form:

> 'It is *logically* possible that: *p* because *q*', which is equivalent to 'It is *causally* possible that: *p* because *q* and initial conditions *c*', where we may not know which conditions these are, nor whether they hold for the given empirical target.

This classification is similar to Gräbner's (2018) proposal, where his 'full explanations' correspond to what I call HAEs, his 'partial explanations' to causal HPEs, and his 'potential explanations' include causal HPEs and ep-op HPEs.[12] The notion of ep-op HPEs is closely related to what Ylikoski and Aydinonat (2014) call 'causal mechanism schemes', which 'do not directly explain any particular empirical fact' but 'address only simplified theoretical explananda' (Ylikoski and Aydinonat, 2014, 27). By calling such HPEs epistemically opaque, we highlight the indeterminate nature of the represented target phenomenon.

---

notion across literature. My approach here is in line with Verreault-Julien (2018) in terms of assigning a broad meaning to HPEs.

[9] I consider causal explanations because they are typically discussed in the context of the modelling in social sciences, see e.g. Alexandrova (2008), Northcott and Alexandrova (2015), Reiss (2012).

[10] The latter case captures counterfactual scenarios, while the former one captures potential scenarios, which may be actual or counterfactual.

[11] For a detailed account of epistemically opaque HPEs, see Šešelja et al. (2022).

[12] For a more general discussions on different types of explanations obtained by means of models see Bokulich (2017), Lawler and Sullivan (2020).

The 'initial conditions' mentioned in the classification above stand for various contextual factors that must be satisfied for a particular regularity to hold. In case of ABMs of science, this may include for instance the size of the community, the nature of interaction among scientists, the nature of decision-making of scientists concerning theories they want to pursue, etc. Such factors are implicitly or explicitly assumed in the given model.

Note also that in the above, 'knowledge' is used in a colloquial rather than the strictly epistemological sense, and it could be replaced with 'having a justified belief'. The idea is that the conditions constraining a particular explanatory relationship are established via a suitable scientific method, in which case we have a good reason to believe which conditions these are or whether they hold for a given empirical target.

Most models fall somewhere in-between the above categories. Depending on the epistemic status of the initial conditions *c* (whether we are able to specify which ones they are and whether they hold for the empirical target in question), an ABM will be closer to one rather than another type. This is determined by the process of model validation, to which we shall turn now.

4 Verification and validation of ABMs

The main reason for running simulations of a scientific inquiry is to examine the impact of certain factors on the collective goals of research, such as efficiency, which would be difficult to estimate by analytical methods or by the means of qualitative analysis. This means that the results of an adequate ABM should not be merely obvious consequences of the underlying assumptions (Lazer and Friedman, 2007; Pöyhönen and Kuorikoski, 2016), because that would make the entire process of modelling superfluous. This, however, means that the link between the model and its purported target need not be obvious. In particular, when a highly idealised model is first proposed, the results it delivers may come with a degree of epistemic opacity in the sense that we do not understand the conditions under which the established causal dependency holds. To remove this veil of opacity, we need to turn to the validation procedures.[13]

Justification of models and their representational properties is conducted via two closely related processes: verification and validation. While verification is a method of evaluating the accuracy of the program of a given ABM based on its conceptual design, validation is the process of evaluation of links between the model and its purported target (e.g., Cooley and Solano, 2011; Gräbner, 2018). Irrespective

[13] This corresponds to what Bokulich (2011) calls a 'justificatory step' in establishing explanations obtained by a model, i.e., the domain of its applicability.

of the purpose for which the model was built, it always requires some degree of verification to ensure that its simulation code does not suffer from bugs and other unintended issues. In short, that it corresponds to the modeller's conceptual idea. The type of required validation, however, directly depends on the purpose of the model and its intended target. In particular, examination of whether the model represents a logical possibility, a real-world possibility, or a real-world phenomenon requires different validation procedures.

Clearly, showing that a model represents a logical possibility will be the least demanding of the procedures alluded to above. All that needs to be shown is that there is a plausible interpretation of the model such that the inference 'it is logically possible that: *p* because *q*' is warranted. For instance, if an ABM is supposed to represent the impact of a certain division of cognitive labour among scientists on the success of their inquiry, we only need to show that we can plausibly interpret the model as representing scientific research and the division of cognitive labour among scientists. At the same time, we need not know under what particular conditions of inquiry (e.g., for how large a community, under what communication structure, under what research behaviour of scientists, etc.) the observed regularity (here between a specific division of labour and a particular measure of success) holds.

This need not, however, be the only goal we are interested in. Even in the case of abstract, highly idealised models, we are often after more than a mere logical possibility. For instance, we may be interested in showing that a typical case of scientific inquiry within a certain domain of study is at least 'susceptible' towards a particular regularity.[14] In other words, we may be interested in causal scenarios which are possible under a set of conditions typical of inquiries within a given scientific domain. To achieve this, we need a model that provides a causal HPE. To go back to the example above, it would mean showing the impact of a specific division of labour on the success of inquiry under a set of conditions typical for research in a particular scientific domain.

Since the difference between causal HPE and ep-op HPE rests in the epistemic status of the initial conditions under which the observed regularity holds, the better we can specify such conditions, the more we are able to move away from an ep-op HPE and towards a causal HPE. This is where various validation procedures enter the stage. On the one hand, their purpose is to help us determine the conditions in the model world under which the results of simulations remain stable. On the other hand, validation helps us to relate these conditions to empirical phenomena. The former is the task of robustness analysis and the latter of an empirical embedding of the model.

[14] For example, Nguyen (2019) takes the Schelling model as licensing the claim: 'A city whose residents have weak preferences regarding the skin colour of their neighbours has a susceptibility towards global segregation.' He does not tell us, however, in virtue of what exactly such susceptibility can be considered warranted.

4.1 Robustness analysis

As the name suggests, robustness analysis is a method of examining the robustness, or stability, of results of a particular model under changes in its assumptions. Depending on the kind of assumptions we focus on, we can distinguish between two types of analyses:

a) *Sensitivity analysis* is a method of examining the robustness of results under changes in the values of parameters in the model (Thiele et al., 2014).[15] This analysis is used to determine the scope of parameters within which the results of a simulation remain stable.

b) *Derivational robustness analysis* is a method of examining the robustness of results under changes in the (idealising) assumptions of the model.[16] This is especially important in the case of highly idealised models, where it is usually difficult to assess whether idealisations impact the results or not. One way of conducting a derivational robustness analysis is by using a family of ABMs to gradually vary the assumptions of the initial model and examine how such changes impact the results (Aydinonat et al., 2020). Another option is to use structurally different models aimed at representing the same target phenomenon: this approach can help reveal the impact of implicit assumptions and idealisations.

While robustness analysis can help us to better understand the ABM in question, it is typically insufficient as a method of specifying the empirical conditions under which particular results hold (see, e.g., Houkes and Vaesen, 2012). For instance, if the analysis shows that the results are relatively stable, we may still have insufficient evidence to claim that they are representative of a given empirical target. Perhaps a specific assumption in the model whose impact has not yet been examined could be making all the difference. Or it could be the case that the empirical target is best represented in terms of very specific parameter values, which have not been carefully examined by robustness tests. To amend this problem, robustness analysis needs to be supplemented with, and guided by, an empirical embedding of the model.

[15] Gräbner (2018) considers sensitivity analysis a part of verification rather than validation, because its purpose is to explore the results, rather than link them to a specific target. This view, however, disregards the fact that sensitivity analysis can be informative in this sense as well. For example, if it turns out a particular result occurs only under a small portion of the parameter space, this would pose an additional requirement on examining whether these parameters correspond to any empirical circumstances.

[16] Derivational robustness construed this way includes both 'structural robustness' and 'representational robustness' as defined by Weisberg and Reisman (2008), where the former stands for stability of the results under changes in the causal structure of the modelled system, and the latter for stability of the results under changes in the representational framework of the model. For discussions on derivational and representational robustness, see Woodward (2006), Ylikoski and Aydinonat (2014), Lehtinen (2017), Railsback and Grimm (2011, 302–306), and Kuhlmann (2021).

4.2 Empirical embedding and model validation

As mentioned above, the robustness analysis can be guided towards an examination of those assumptions that correspond to the intended empirical target. This allows us to check whether the causal dependency inferred from the model holds under assumptions which are empirically relevant. But how does one make sure the relevant assumptions are well embedded and indeed correspond to the relevant empirical phenomena? This is done via different strategies jointly known as empirical validation of ABMs.[17] Following Gräbner (2018), I list some of the most relevant procedures.

a) *Process validation* concerns the question of how well mechanisms represented in the model reflect our empirical knowledge about them (Gräbner, 2018). To this end, the strategy of enhancing the theoretical realism of the model by information based on our knowledge from sociology and the philosophy of science can be helpful (Casini and Manzo, 2016; Šešelja, 2021). For instance, exchange of information among scientists has been typically represented as a simple sharing of results of scientific studies (e.g., Grim et al., 2013; Weisberg and Muldoon, 2009; Zollman, 2010), but qualitative philosophical accounts of scientific communication often emphasise critical interaction (e.g., Longino, 2002, Chang, 2012). For this reason, inclusion of this aspect in ABMs of science when examining the robustness of previously obtained results may be one way of conducting their process validation (e.g., Borg et al., 2018; Frey and Šešelja, 2020).

b) *Input validation* concerns the question of whether the exogenous inputs for the model are empirically meaningful and appropriate for the purpose at hand (Tesfatsion, 2017). This may include behavioural assumptions ascribed to the agents, the initial conditions, parameter values, etc. (Fagiolo et al., 2019). If parameters in the model are adjusted so as to reflect or include concrete numerical information, we say a model is 'empirically calibrated' (Boero and Squazzoni, 2005). In the case of ABMs of science, this would mean for example adjusting the number of agents in a model according to the size of a particular scientific community or representation of social networks in the model based on bibliometric data (Martini and Pinto, 2016; Perović et al., 2016; Thicke, 2019).

c) *Descriptive and predictive output validation* concern the question to what extent the output of the model replicates existing knowledge about the target and whether it can predict its future states (Gräbner, 2018; Tesfatsion, 2017; Thicke, 2019). For instance, if a model aims at representing a certain episode from the history of science, then under specific initial conditions the macrobehaviour of simulated agents should correspond to our historical knowledge of the case study in question.

[17] Literature on this topic is plentiful, see, e.g., Arnold (2019), Beisbart and Saam (2018), Boero and Squazzoni (2005), Casini and Manzo (2016), Gräbner (2018), Guerini and Moneta (2017), Richiardi et al. (2006), Tesfatsion (2017), Thicke (2020).

All in all, validation of ABMs is essential for determining the details of targets they represent. In particular, validation supplements and guides the robustness analysis in determining the conditions under which the causal dependency identified via the model holds. By following the above validation strategies, we can move the explanation based on a particular model from 'epistemic opaqueness' to a causal HPE (or to a HAE). In the following section, I illustrate this point with a class of ABMs of scientific interaction.

5 ABMs of scientific interaction: zooming in on the target

In this section, I look into a class of ABMs which were developed to represent the effects of scientific interaction on the efficiency of inquiry. The main question these models aim to address is how different degrees of connectedness across a given scientific community impact the efficiency of knowledge acquisition. While at first sight, a high degree of interaction would seem purely beneficial, simulations have shown that this need not always be the case. For instance, if misleading information spreads quickly through the scientific community, scientists may collectively end up choosing a wrong theory.

To understand the root of this problem, it is useful to clarify the trade-off between 'exploration' and 'exploitation', to which it is closely related. The relationship between exploration (search for new possibilities) and exploitation (the use of existing options) has long been studied in theories of formal learning, organisational sciences, etc. (March, 1991). It is easy to see that a similar trade-off may take place in the context of scientific inquiry: given a particular scientific problem, one can either explore novel ideas and hope to find solutions which are better than the existing ones, or stick with the currently available hypotheses and use those instead. Depending on the difficulty of the problem, different strategies of balancing between exploration and exploitation are more suitable: for instance, if a solution to a problem is hard to find, scientists may need to invest their resources in exploration before focusing on exploiting existing ideas.

Simulations of scientific interaction were inspired by the idea that different communication networks among scientists, characterised by varying degrees of connectedness (see Figure 2), may have a different impact on the balance between exploration and exploitation. In particular, if an initially misleading idea is shared too quickly through the community, scientists may lock in on it and prematurely abandon their search for better solutions. Alternatively, if the information flow is slow and sparse, important insights gained by some scientists, which could lead to an optimal solution, may remain undetected by the rest of the community for a long time.

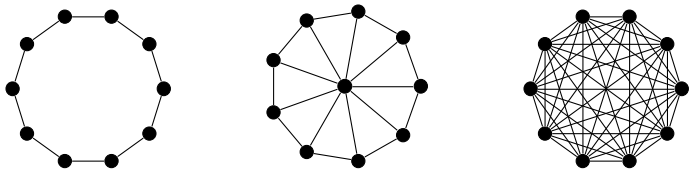

Figure 2: Three types of communication networks, representing an increasing degree of connectedness: a cycle, a wheel, and a complete graph. The nodes in each graph stand for scientists, while edges between the nodes stand for transmission of information between two scientists.

In what follows, I will look at a class of ABMs of scientific interaction starting with the pioneering work by Kevin Zollman. After suggesting an epistemically opaque how-possibly explanation (ep-op HPE) that can be drawn from his models, I proceed to examine how subsequent research allowed for specification of further conditions under which the observed regularity holds.

5.1 Scientific interaction and bandit problems

A set of ABMs developed by Zollman (2007, 2010, 2013) is based on the idea that scientific interaction can be studied in terms of 'bandit problems'. Bandit problems, well-known in economics and statistics, are a prime example of the exploration–exploitation trade-off. They concern a situation in which a gambler, or a group of gamblers, is confronted with multiple slot machines ('bandits'), which have different probabilities of success. While gamblers aim to maximise their overall reward, it is not immediately clear how long they should test each available machine and at which point they should stick with one that seems to give the highest payoff. If we further suppose that gamblers can share information among themselves and that each gambler sticks to the machine that seems to give the highest reward, we can ask: Which communication network will increase their chance to identify the machine with the highest payoff?

Zollman starts with the idea that this type of uncertainty is similar to one which scientists find themselves in when confronted with multiple rival hypotheses. Using a framework developed by Bala and Goyal (1998), he investigates which types of communication networks increase the chance that a scientific community, confronted with two rival hypotheses, successfully identifies the better of the two.

At the beginning of the simulation,[18] scientists – represented as Bayesian reasoners – are assigned random prior probabilities for two rival hypotheses, each of which has a designated objective probability of success. Agents always choose to pursue a theory which they believe to be better. During the simulation, they update their beliefs based on their own findings and the information they receive from their

[18] I am describing Zollman's (2010) model, which is a generalised version of his 2007 proposal.

neighbours within a particular social network. Zollman examines three kinds of social networks from Figure 2. Scientists are successful if they manage to converge on the objectively better hypothesis (i.e., one that has a higher objective probability of success).

His results suggest that a high degree of interaction can be harmful. Because the initial findings about the hypotheses may be misleading, when scientists are linked via a complete graph the misleading information will spread quickly throughout the community. Consequently, the entire community may prematurely abandon the objectively better hypothesis.

Zollman also observes that if scientists start with extreme prior values, representing agents who stick to their hypotheses, the misleading information will not affect them early on. In fact, the complete graph is in such scenario more successful than the cycle.[19]

Altogether, the simulation results in the following ep-op HPE:

> **(High-inf)** It is logically possible that a scientific community prematurely abandons the better of two rival hypotheses because of a high degree of information flow among the scientists.

To turn a *High-inf* into a causal HPE from which we could make inferences about real-world possibilities, we need to specify the conditions under which a particular regularity holds. While Zollman provides one such condition, namely the absence of extreme priors, subsequent research has examined some additional factors.

5.2 The context of difficult inquiry

A number of related studies had shown that the main domain of application of Zollman's results is the context of a difficult inquiry. I take a brief look at these results and classify them according to the type of validation procedure they support.

Sensitivity analysis. Rosenstock et al. (2017) conducted a sensitivity analysis of Zollman's findings and showed that the 'Zollman effect' – the superior performance of the cycle versus the complete graph – holds only for a small part of the relevant parameter space. In particular, they show that the result obtains when the two relevant hypotheses are similar in terms of their objective probability of success, the population size is small, and the amount of data collected by scientists on each round is likewise small. The authors conclude that these factors are characteristic of *difficult learning*, because scientists either have a hard time distinguishing between the rival hypotheses or their data is sparse. Such conditions make

[19] This result is obtained by stopping the simulation after a certain number of rounds. Given sufficient time, agents in all networks end up on the correct hypothesis.

it easier for misleading information to propagate through the community and sway it to the wrong hypothesis.

All in all, the results of sensitivity analysis restrict the application domain of *High-inf* to the context of difficult inquiry.

Derivational robustness. Restriction of the application domain to the context of difficult inquiry finds further support in results obtained by some structurally different ABMs. First, the ABMs by Lazer and Friedman (2007), which were developed in organisational sciences, arrived at a similar conclusion. Their model is designed to study the problem-solving performance of agents linked via different social networks using a multidimensional epistemic landscape. The authors observe that in complex tasks that require a problem-solving capacity to extend over a longer period of time, highly connected networks perform worse than the less connected ones. Similar to what happens in Zollman's model, highly connected groups quickly converge on a single approach, thus failing to preserve the diversity of ideas needed to solve complex tasks.

Results supporting *High-inf* have also been obtained with subsequent ABMs based on epistemic landscapes (e.g., Grim, 2009; Grim et al., 2013, Derex et al., 2018), which suggests their derivational robustness (although see below).

Empirical output validation. Finally, the output of these models was reproduced by some empirical studies. For example, Mason et al. (2008) as well as Derex and Boyd (2016) conducted computer-based experiments in which participants linked via different communication networks were confronted with certain problem-solving situations. Both studies concluded that less interconnected groups outperform the more connected ones because they are able to preserve diversity and explore the space of possible solutions to a higher degree.

While all of these findings support *High-inf* under the conditions of difficult learning, we ought to be cautious with their extrapolation to actual scientific inquiries. One thing to note is that all of the above mentioned studies are based on the assumption – integral to both the simulations and the experimental setup of empirical studies – that there is a trade-off between exploitation and exploration. But it should be noted that neither is actual scientific inquiry necessarily based on this trade-off, nor do results obtain once the trade-off assumption is relaxed.

5.3 Relaxing the exploration/exploitation trade-off

When scientists pursue a theory, it is not uncommon that along the way they acquire information relevant to the assessment of a rival theory. For example, scientists may detect some explanatory anomalies in their current theory (e.g., evidence that cannot be accounted for by that theory) that could be explained by the rival theory. As a result, research into the former (exploitation) could inspire and lead to research on the latter (exploration).

These considerations inspired ABMs and empirical studies that relaxed the exploration/exploitation trade-off. Here, I review some examples.

Derivational robustness in view of exploratory agents. Kummerfeld and Zollman (2016) developed an ABM of scientific interaction based on an analogy with bandit problems, but this time allowing agents who pursue one hypothesis to also occasionally acquire information about a rival hypothesis. Their results show that higher levels of exploration by agents go hand in hand with benefits of increased connectivity among them.

The positive impact of high levels of interaction has been observed also in a structurally different model: argumentation-based ABM (ArgABM) (Borg et al., 2019, 2017, 2018). ArgABM aims at capturing the argumentative dynamics underlying a scientific inquiry. The model employs an 'argumentative landscape' representing rival research programmes or theories in a given domain which scientists gradually explore. Each theory consists of 'arguments', which stand for studies supporting a particular theory. These arguments can be challenged ('attacked') by studies belonging to rival research programmes or defended by further arguments developed within the same programme. In this way, the argumentative landscape allows for the representation of both false positives (acceptance of a false hypothesis) and false negatives (rejection of a true hypothesis). The success of inquiry is measured in terms scientists converging on the theory that is predefined as fully defensible within the landscape (initially unknown to the agents).

The results of ArgABM indicate that a high degree of interaction among scientists is beneficial. The more connected agents are, the better their chances of converging on the best theory, and this holds under a variety of conditions of inquiry.

The main reason ArgABM delivers this result lies in the following two modelling assumptions. First, when agents explore a theory, they also gain information about rival theories in the form of argumentative attacks or defences of own theory. For instance, by finding an argument in my theory that attacks the rival

theory, I identify a potential problem in the latter. Alternatively, if I encounter an attack on my own theory, I will learn the argument from the rival theory (this could represent a scenario in which proponents of the rival theory publish a study showing they are able to explain certain phenomena which our theory cannot explain that well). As a result, exploitation includes a degree of exploration.

Second, to accurately evaluate a theory (e.g., in terms of the number of 'anomalies' represented as attacked and undefended arguments in a theory, see Borg et al., 2019), agents need a sufficiently detailed knowledge of the argumentative landscape. If a scientist knows only a part of the landscape, she may assess a particular theory as unproblematic, while in fact she has not learned about its problematic parts. This corresponds to a scenario in which scientists, having read a few studies in favour of a particular research programme, conclude that the programme is feasible, but they failed to read other studies, which show that results presented in the former ones could not be replicated or are based on a methodological error. As a result, less connected groups will suffer from greater information losses, making it more likely that their assessment of a particular theory is inaccurate.

Empirical output validation. In contrast to the previously mentioned empirical studies, an experiment run by Mason and Watts (2012) resulted in the conclusion that a higher degree of connectivity is actually rewarding. Unlike the former experiments, this study is based on a relaxed assumption about the exploration/exploitation trade-off. Exploitation of existing ideas does not necessarily restrict participants to the local maxima. Instead, they have the option of going on to individually search for better solutions.

In sum, several studies that relaxed the assumption about the trade-off between exploration and exploitation failed to replicate *High-inf*, thus pointing to limitations of its application. Additionally, the ArgABM highlighted the negative aspect of information loss that can take place in loosely connected communities.

5.4 Alternative mechanisms of diversity

Derivational robustness under the assumption of cautious agents. Frey and Šešelja (2020) have conducted an additional derivational robustness analysis of Zollman's (2010) model by enhancing it with a number of assumptions characteristic of a difficult inquiry.[20] That study is therefore also a contribution to the *process validation* mentioned in Section 4, while more specifically, it focuses on the robustness of results once the process of difficult inquiry is captured in terms of empirically relevant assumptions.

[20] The code of their model, available at https://github.com/daimpi/SocNetABM/tree/ RobIdeal, also includes Zollman's ABM as a nested variant and thus provides an easily accessible tool for its verification.

The most important finding of those simulations is that even in the context of difficult inquiry, a high degree of information flow is not necessarily harmful. On the contrary, more connected networks may outperform the less connected ones. In particular, if diversity is generated in some other way than by the means of network structure, a high information flow will not have a negative impact on the efficiency of the group. For instance, if scientists are equipped with a dose of caution, or 'rational inertia', when deciding whether they should abandon their current theory and start pursuing a rival one, the cycle is no longer superior to the complete graph.

Moreover, addition of the assumption that agents interact critically does not on its own help the complete graph to catch up with the cycle: for that to happen, scientists must be cautious in their decision-making (for instance by displaying a degree of resistance against changing the theory they had endorsed).

All in all, these results further specify conditions under which the *High-inf* holds.

5.5 From ep-op HPE to causal HPE

To sum up, the studies reviewed above suggest that the explanation obtained from Zollman's original model can be expressed as follows:

(High-inf-causal) It is causally possible that a scientific community prematurely abandons the better of two rival hypotheses due to a high degree of information flow among scientists under the following conditions:

- that the inquiry is difficult
- theoretical diversity is not generated in some other way (e.g., by scientists having extreme priors or a tendency to stick to their hypotheses)
- that pursuit of one hypothesis does not allow for insights into its rivals (i.e., there is a strict exploration/exploitation trade-off)
- potentially some additional assumptions.

While in the original model, we could only draw an ep-op HPE without a clear application domain, subsequent studies allowed us to zoom in on the target that the model actually represents and for which the observed causal mechanism appears to hold. Of course, further studies may reveal that additional specifications are needed or that some of the existing ones ought to be revised.

The preceding discussion also illustrates that the difficulty of extrapolating findings from a model to an empirical application domain holds not only for ABMs but also for empirical experimental studies.

6 Conclusion

In this paper, I explored the epistemic benefits of running computer simulations in the philosophy of science and the kinds of inferences one can draw from them. I have argued that models can represent (i) logical possibilities, (ii) real-world possibilities, or (iii) real-world phenomena, where each category comes with specific explanatory features. By using strategies of verification and validation, we can identify the class to which a particular ABM belongs. While abstract, highly idealised models prima facie allow only for the inference of a causal possibility under unknown circumstances, the process of validation by the means of other ABMs as well as empirical studies can help reveal these conditions.

In conclusion, let me make a few general points. First, highly idealised ABMs of science should be appreciated even under conditions of a minimal degree of verification and validation required for obtaining ep-op HPE. In this form, they can assume a variety of epistemic functions, ranging from providing conjectures about scientific inquiry and starting a new family of models all the way to contributing to the validation of other ABMs. Second, the development of new ABMs and their subsequent validation is best considered in terms of broader inquiries consisting of classes of ABMs, but also empirical studies targeting the same phenomenon. Third, there is no reason to see the highly idealised nature of ABMs of science as their drawback. As long as the model is subjected to an adequate process of verification and validation with respect to its purported aim and target, it can be an important step forward in our understanding of scientific inquiry.

Acknowledgements. I am very grateful to Daniel Frey, Paul Hoyningen Huene, Christian Straßer, Emily Sullivan, Leonid Tiokhin, Philippe Verreault-Julien, and the Philosophy & Ethics Group at TU Eindhoven for fruitful discussions that helped me formulate a number of points in this paper. Research on this this paper is based was funded by Irène Curie Fellowship of TU Eindhoven and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 426833574.

References

Alexander, Jason McKenzie, Johannes Himmelreich, and Christopher Thompson (2015). "Epistemic landscapes, optimal search, and the division of cognitive labor". In: *Philosophy of Science* 82.3, pp. 424–453.

Alexandrova, Anna (2008). "Making models count". In: *Philosophy of Science* 75.3, pp. 383–404.

Arnold, Eckhart (2019). “Validation of Computer Simulations from a Kuhnian Perspective”. In: *Computer Simulation Validation – Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*. Cham: Springer.

Avin, Shahar (2019). “Centralized funding and epistemic exploration”. In: *The British Journal for the Philosophy of Science* 70.3, pp. 629–656.

Axelrod, Robert (1984). *The evolution of cooperation*. Basic Books.

Axelrod, Robert (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*. Vol. 3. Princeton University Press.

Axelrod, Robert and William Donald Hamilton (1981). “The evolution of cooperation”. In: *Science* 211.4489, pp. 1390–1396.

Aydinonat, N. Emrah, Samuli Reijula, and Petri Ylikoski (2020). “Argumentative Landscapes: The Functions of Models in Social Epistemology”. In: *Synthese* (forthcoming).

Bala, Venkatesh and Sanjeev Goyal (1998). “Learning from neighbours”. In: *The review of economic studies* 65.3, pp. 595–621.

Balietti, Stefano, Michael Mäs, and Dirk Helbing (2015). “On Disciplinary Fragmentation and Scientific Progress”. In: *PloS one* 10.3.

Bedessem, Baptiste (2019). “The division of cognitive labor: Two missing dimensions of the debate”. In: *European Journal for Philosophy of Science* 9.1, pp. 1–16.

Beisbart, Claus and Nicole J Saam, eds. (2018). *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*. Cham: Springer.

Boero, Riccardo and Flaminio Squazzoni (2005). “Does empirical embeddedness matter? Methodological issues on agent-based models for analytical social science”. In: *Journal of artificial societies and social simulation* 8.4.

Bokulich, Alisa (2011). “How scientific models can explain”. In: *Synthese* 180.1, pp. 33–45.

Bokulich, Alisa (2014). “How the tiger bush got its stripes:‘How possibly’vs.‘how actually’model explanations”. In: *The Monist* 97.3, pp. 321–338.

Bokulich, Alisa (2017). “Models and explanation”. In: *Springer handbook of model-based science*. Springer: Cham, pp. 103–118.

Bolinska, Agnes (2013). “Epistemic representation, informativeness and the aim of faithful representation”. In: *Synthese* 190.2, pp. 219–234.

Borg, AnneMarie, Daniel Frey, Dunja Šešelja, and Christian Straßer (2019). "Theory-choice, transient diversity and the efficiency of scientific inquiry". In: *European Journal for Philosophy of Science*. http://doi.org/10.1007/s13194-019-0249-5.

Borg, AnneMarie, Daniel Frey, Dunja Šešelja, and Christian Straßer (2017). "Examining Network Effects in an Argumentative Agent-Based Model of Scientific Inquiry". In: *Logic, Rationality, and Interaction: 6th International Workshop, LORI 2017, Sapporo, Japan, September 11-14, 2017, Proceedings*. Ed. by Alexandru Baltag, Jeremy Seligman, and Tomoyuki Yamada. Berlin–Heidelberg: Springer, pp. 391–406.

Borg, AnneMarie, Daniel Frey, Dunja Šešelja, and Christian Straßer (2018). "Epistemic effects of scientific interaction: approaching the question with an argumentative agent-based model". In: *Historical Social Research* 43.1, pp. 285–309.

Bruch, Elizabeth and Jon Atwell (2015). "Agent-based models in empirical social research". In: *Sociological methods & research* 44.2, pp. 186–221.

Casini, Lorenzo and Gianluca Manzo (2016). "Agent-based models and causality: a methodological appraisal". In: *(The IAS Working Paper Series). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva133332*.

Chang, Hasok (2012). *Is Water H2O? Evidence, Pluralism and Realism*. Cham: Springer.

Cooley, Philip and Eric Solano (2011). "Agent-based model (ABM) validation considerations". In: *Proceedings of the Third International Conference on Advances in System Simulation (SIMUL 2011)*, pp. 134–139.

Currie, Adrian and Shahar Avin (2018). "Method Pluralism, Method Mismatch & Method Bias". In: *Philosopher's Imprint* 19.13, pp. 1–22.

Derex, Maxime and Robert Boyd (2016). "Partial connectivity increases cultural accumulation within groups". In: *Proceedings of the National Academy of Sciences* 113.11, pp. 2982–2987.

Derex, Maxime, Charles Perreault, and Robert Boyd (2018). "Divide and conquer: Intermediate levels of population fragmentation maximize cultural accumulation". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1743, p. 20170062.

Douven, Igor (2010). "Simulating peer disagreements". In: *Studies in History and Philosophy of Science Part A* 41.2, pp. 148–157.

Dray, William H (1957). "Laws and explanation in history". 3rd edition. Oxford: Oxford University Press.

Edmonds, Bruce and Scott Moss (2004). "From KISS to KIDS–an 'antisimplistic'modelling approach". In: *International workshop on multiagent systems and agent-based simulation*. Springer, pp. 130–144.

Edmonds, Bruce, Christophe le Page, Volker Grimm, Cristina Montanola, Paul Ormerod, Hilbert Root, and Flaminio Squazzoni1 (2018). "Different Modelling Purposes". In: forthcoming.

Epstein, Joshua M (2006). *Generative social science: Studies in agent-based computational modelling*. Princeton, NJ: Princeton University Press.

Epstein, Joshua M (2009). "Modelling to contain pandemics". In: *Nature* 460.7256, pp. 687– 687.

Epstein, Joshua M and Robert Axtell (1996). *Growing artificial societies: social science from the bottom up*. Washington, DC: Brookings Institution Press.

Fagiolo, Giorgio, Mattia Guerini, Francesco Lamperti, Alessio Moneta, and Andrea Roventini (2019). "Validation of agent-based models in economics and finance". In: *Computer Simulation Validation*. Cham: Springer, pp. 763–787.

Forber, Patrick (2010). "Confirmation and explaining how possible". In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 41.1, pp. 32–40.

Forber, Patrick (2012). "Conjecture and explanation: A reply to Reydon". In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43.1, pp. 298–301.

Frey, Daniel and Dunja Šešelja (2018). "What is the Epistemic Function of Highly Idealized Agent-Based Models of Scientific Inquiry?" In: *Philosophy of the Social Sciences* https://doi.org/10.1177/0048393118767085.

Frey, Daniel and Dunja Šešelja (2020). "Robustness and Idealization in Agent-Based Models of Scientific Interaction". In: *British Journal for the Philosophy of Science* 71, pp. 1411–1437. url: https://doi.org/10.1093/bjps/axy039.

Fumagalli, Roberto (2016). "Why we cannot learn from minimal models". In: *Erkenntnis* 81.3, pp. 433–455.

Gilbert, Nigel (1997). "A simulation of the structure of academic science". In: *Sociological Research Online* 2.2, pp. 1–15.

Gilbert, Nigel and Klaus Troitzsch (2005). *Simulation for the social scientist*. London: McGraw-Hill Education.

Goldman, Alvin I and Moshe Shaked (1991). "An economic model of scientific activity and truth acquisition". In: *Philosophical Studies* 63.1, pp. 31–55.

Gräbner, Claudius (2018). "How to Relate Models to Reality? An Epistemological Framework for the Validation and Verification of Computational Models". In: *Journal of Artificial Societies and Social Simulation* 21.3, p. 8. issn: 1460-7425. doi: 10.18564/jasss.3772. url: http: //jasss.soc.surrey.ac.uk/21/3/8.html.

Grim, Patrick (2009). "Threshold Phenomena in Epistemic Networks." In: *AAAI Fall Symposium: Complex Adaptive Systems and the Threshold Effect*, pp. 53–60.

Grim, Patrick, Horace Paul St, Gary Mar, Paul Saint Denis, and Paul St Denis (1998). *The philosophical computer: Exploratory essays in philosophical computer modelling*. Vol. 1. Cambridge, MA: MIT Press.

Grim, Patrick and Daniel Singer (2020). “Computational Philosophy”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2020. Metaphysics Research Lab, Stanford University.

Grim, Patrick, Daniel J Singer, Steven Fisher, Aaron Bramson, William J Berger, Christopher Reade, Carissa Flocken, and Adam Sales (2013). “Scientific networks on data landscapes: question difficulty, epistemic success, and convergence”. In: *Episteme* 10.04, pp. 441–464.

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

Guerini, Mattia and Alessio Moneta (2017). “A method for agent-based models validation”. In: *Journal of Economic Dynamics and Control* 82, pp. 125–141.

Harnagel, Audrey (2018). “A Mid-Level Approach to Modelling Scientific Communities”. In: *Studies in History and Philosophy of Science*. https://doi.org/10.1016/j.shpsa.2018.12.010.

Hegselmann, Rainer (2017). “Thomas C. Schelling and James M. Sakoda: The intellectual, technical, and social history of a model”. In: *Journal of Artificial Societies and Social Simulation* 20.3.

Hegselmann, Rainer and Ulrich Krause (2002). “Opinion dynamics and bounded confidence models, analysis, and simulation”. In: *Journal of artificial societies and social simulation* 5.3.

Hegselmann, Rainer and Ulrich Krause (2005). “Opinion dynamics driven by various ways of averaging”. In: *Computational Economics* 25.4, pp. 381–405.

Hegselmann, Rainer and Ulrich Krause (2006). “Truth and cognitive division of labor: First steps towards a computer aided social epistemology”. In: *Journal of Artificial Societies and Social Simulation* 9.3, p. 10.

Hempel, Carl (1965). *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York: Free Press.

Holman, Bennett and Justin Bruner (2015). “The problem of intransigently biased agents”. In: *Philosophy of Science* 82.5, pp. 956–968.

Holman, Bennett and Justin Bruner (2017). “Experimentation by industrial selection”. In: *Philosophy of Science* 84.5, pp. 1008–1019.

Houkes, Wybo and Krist Vaesen (2012). “Robust! Handle with care”. In: *Philosophy of Science* 79.3, pp. 345–364.

Hoyningen-Huene, Paul (2020). “The Logic of Explanation by Abstract Models”. In: forthcoming.

Imbert, Cyrille (2017). “Computer simulations and computational models in science”. In: *Springer handbook of model-based science*. Cham: Springer, pp. 735– 781.

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

Kitcher, Philip (1993). *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford: Oxford University Press.

Klein, Dominik, Johannes Marx, and Kai Fischbach (2018). "Agent-based modelling in social science, history, and philosophy. An introduction". In: *Historical Social Research/Historische Sozialforschung* 43.1 (163, pp. 7– 27.

Kuhlmann, Meinard (2021). "On the Exploratory Function of Agent-Based Modelling". In: *Perspectives on Science* 29.4, pp. 510–536.

Kummerfeld, Erich and Kevin JS Zollman (2016). "Conservatism and the scientific state of nature". In: *The British Journal for the Philosophy of Science* 67.4, pp. 1057–1076.

Lawler, Insa and Emily Sullivan (2020). "Model Explanation versus Model-Induced Explanation". In: *Foundations of Science* 26, pp. 1049–1074.

Lazer, David and Allan Friedman (2007). "The network structure of exploration and exploitation". In: *Administrative science quarterly* 52.4, pp. 667–694.

Lehtinen, Aki (2017). "Derivational robustness and indirect confirmation". In: *Erkenntnis*, pp. 1–38.

Longino, Helen E (2002). "Science and the common good: Thoughts on Philip Kitcher's Science, Truth, and Democracy". In: *Philosophy of Science* 69.4, pp. 560–568.

Mäki, Uskali (2005). "Economic epistemology: Hopes and horrors". In: *Episteme* 1.03, pp. 211–222.

March, James G (1991). "Exploration and exploitation in organizational learning". In: *Organization science* 2.1, pp. 71–87.

Martini, Carlo and Manuela Fernández Pinto (2016). "Modelling the social organization of science". In: *European Journal for Philosophy of Science*, pp. 1–18.

Mason, Winter and Duncan J Watts (2012). "Collaborative learning in networks". In: *Proceedings of the National Academy of Sciences* 109.3, pp. 764–769.

Mason, Winter A, Andy Jones, and Robert L Goldstone (2008). "Propagation of innovations in networked groups." In: *Journal of Experimental Psychology: General* 137.3, p. 422.

Mayo-Wilson, Conor and Kevin JS Zollman (2021). "The computational philosophy: simulation as a core philosophical method". In: *Synthese*, pp. 1–27.

Muldoon, Ryan and Michael Weisberg (2011). "Robustness and idealization in models of cognitive labor". In: *Synthese* 183.2, pp. 161–174.

Nguyen, James (Mar. 2019). "It's Not a Game: Accurate Representation with Toy Models". In: doi: 10.1093/bjps/axz010. eprint: http://oup. prod.sis.lan/bjps/advance-article-pdf/doi/10.1093/bjps/ axz010/28212418/axz010.pdf. url: https://doi.org/10.1093/ bjps/axz010.

Northcott, Robert and Anna Alexandrova (2015). “Prisoner’s Dilemma Doesn’t Explain Much”. In: *The Prisoner’s Dilemma. Classic philosophical arguments.* Ed. by Martin Peterson. Cambridge: Cambridge University Press, pp. 64–84.

O’Connor, Cailin and James Owen Weatherall (2018). “Scientific polarization”. In: *European Journal for Philosophy of Science* 8.3, pp. 855–875.

O’Connor, Cailin and James Owen Weatherall (2019). *The misinformation age: How false beliefs spread*. New Haven, CT: Yale University Press.

Page, Scott E (2018). *The model thinker: what you need to know to make data work for you*. London: Hachette UK.

Payette, Nicolas (2012). “Agent-based models of science”. English. In: *Models of Science Dynamics*. Ed. by Andrea Scharnhorst, Katy Börner, and Peter van den Besselaar. Understanding Complex Systems. Cham: Springer, pp. 127–157.

Perović, Slobodan, Sandro Radovanović, Vlasta Sikimić, and Andrea Berber (2016). “Optimal research team composition: data envelopment analysis of Fermilab experiments”. In: *Scientometrics*, pp. 1–29.

Pinto, Manuela Fernández and Daniel Fernández Pinto (2018). “Epistemic landscapes reloaded: An examination of agent-based models in social epistemology”. In: Historical Social Research/Historische Sozialforschung, 43.1(163), pp. 48–71.

Politi, Vincenzo (2021). “Formal models of the scientific community and the value-ladenness of science”. In: *European Journal for Philosophy of Science*.

Pöyhönen, Samuli (2017). “Value of cognitive diversity in science”. In: *Synthese* 194.11, pp. 4519–4540.

Pöyhönen, Samuli and Jaakko Kuorikoski (2016). “Modelling epistemic communities”. In: *The Routledge Handbook of Social Epistemology*. Ed. by M. Fricker, P. J. Graham, D. Henderson, N. Pedersen, and J. Wyatt. Abingdon: Routledge.

Railsback, Steven F and Volker Grimm (2011). *Agent-based and individualbased modelling: a practical introduction*. Princeton, NJ: Princeton University Press.

Reiss, Julian (2012). “The explanation paradox”. In: *Journal of Economic Methodology* 19.1, pp. 43–62.

Reutlinger, Alexander, Dominik Hangleiter, and Stephan Hartmann (2018). “Understanding (with) Toy Models”. In: *The British Journal for the Philosophy of Science* 69.4, pp. 1069–1099.

Reydon, Thomas AC (2012). “How-possibly explanations as genuine explanations and helpful heuristics: A comment on Forber”. In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43.1, pp. 302–310.

Richiardi, Matteo, Roberto Leombruni, Nicole J. Saam, and Michele Sonnessa (2006). "A Common Protocol for Agent-Based Social Simulation". In: *Journal of Artificial Societies and Social Simulation* 9.1, p. 15. issn: 1460-7425. url: http://jasss.soc.surrey.ac.uk/9/1/15.html.

Rosenstock, Sarita, Cailin O'Connor, and Justin Bruner (2017). "In Epistemic Networks, is Less Really More?" In: *Philosophy of Science* 84.2, pp. 234–252.

Sakoda, James M (1971). "The checkerboard model of social interaction". In: *The Journal of Mathematical Sociology* 1.1, pp. 119–132.

Schelling, Thomas C (1971). "Dynamic models of segregation". In: *Journal of mathematical sociology* 1.2, pp. 143–186.

Schelling, Thomas C (1978). *Micromotives and macrobehavior*. New York, NY: W. W. Norton & Company.

Šešelja, Dunja (2021). "Exploring Scientific Inquiry via Agent-Based Modelling". In: *Perspectives on Science* 29.4. doi: 10.1162/posc_a_00382.

Šešelja, Dunja (2022). "Agent-Based Models of Scientific Interaction". In: Philosophy Compass. Forthcoming.

Šešelja, Dunja, Christian Straßer, and AnneMarie Borg (2020). "Formal Models of Scientific Inquiry in a Social Context: an Introduction". In: *Journal for General Philosophy of Science*. doi: 10.1007/s10838-02009502-w.

Šešelja, Dunja, Philippe Verreault-Julien, and Emily Sullivan (2022). "Epistemically Opaque Explanations". In: *Forthcoming*.

Skyrms, Brian (1990). *The dynamics of rational deliberation*. Cambridge, MA: Harvard University Press.

Skyrms, Brian (1996). *Evolution of the social contract*. Cambridge University Press.

Strevens, Michael (2003). "The role of the priority rule in science". In: *The Journal of philosophy* 100.2, pp. 55–79.

Tesfatsion, Leigh (2017). "Modelling economic systems as locallyconstructive sequential games". In: *Journal of Economic Methodology* 24.4, pp. 384–409.

Thicke, Michael (2020). "Evaluating formal models of science". In: *Journal for General Philosophy of Science* 51, pp. 315–335.

Thiele, Jan C, Winfried Kurth, and Volker Grimm (2014). "Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using NetLogo and R". In: *Journal of Artificial Societies and Social Simulation* 17.3, p. 11.

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

Verreault-Julien, Philippe (2018). “How could models possibly provide howpossibly explanations?” In: *Studies in History and Philosophy of Science Part A*.

Waddell, Paul (2002). “UrbanSim: Modelling urban development for land use, transportation, and environmental planning”. In: *Journal of the American planning association* 68.3, pp. 297–314.

Weatherall, James Owen, Cailin O’Connor, and Justin Bruner (2018). “How to Beat Science and Influence People: Policy Makers and Propaganda in Epistemic Networks”. In: *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axy062.

Weisberg, Michael (2013). *Simulation and similarity: Using models to understand the world*. 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.

Weisberg, Michael and Kenneth Reisman (2008). “The robust Volterra principle”. In: *Philosophy of science* 75.1, pp. 106–131.

Woodward, Jim (2006). “Some varieties of robustness”. In: *Journal of Economic Methodology* 13.2, pp. 219–240.

Ylikoski, Petri (2014). “Agent-based simulation and sociological understanding”. In: *Perspectives on Science* 22.3, pp. 318–335.

Ylikoski, Petri and N Emrah Aydinonat (2014). “Understanding with theoretical models”. In: *Journal of Economic Methodology* 21.1, pp. 19–36.

Zamora Bonilla, Jesús (1999). “The elementary economics of scientific consensus”. In: *Theoria: An International Journal for Theory, History and Foundations of Science*, pp. 461–488.

Zamora Bonilla, Jesús P (2002). “Scientific inference and the pursuit of fame: A contractarian approach”. In: *Philosophy of Science* 69.2, pp. 300–323.

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

Zollman, Kevin JS (2010). “The epistemic benefit of transient diversity”. In: *Erkenntnis* 72.1, pp. 17–35.

Zollman, Kevin JS (2013). “Network epistemology: Communication in epistemic communities”. In: *Philosophy Compass* 8.1, pp. 15–27.