Shaping Science Policy. Instrumental Normativity from Agent-Based Models in Social Epistemology[1]

Thierry Orth

1st Supervisors: Dominik Klein and Erik Stei
2nd Supervisor: Johannes Korbmacher

Philosophy RMA
Department of Philosophy and Religious Studies
Utrecht University

July, 2024

[1] The word count of this thesis amounts to 25,164 words.

Abstract

Agent-based models are models that simulate individuals behaviors with the aim of analysing, evaluating and understanding relationships between micro-level behaviors and macro-level phenomena of complex systems. In recent years, the agent-based approach has found effective entry into social epistemology in the form of so-called SOSR models, which are used to examine relationships between socio-epistemic systems and their epistemic outcomes. An overarching aim of social epistemologists is to use these models to provide normative suggestions for redesigning, restructuring, or reorganising real-world socio-epistemic systems. This thesis defends an instrumentalist view that accounts for this aim in terms of model-based instrumental norms: norms that prescribe socio-epistemic systems as means for achieving epistemic outcomes, which are taken as ends that we have reason to pursue. In laying out this view, we identify three factors relevant to justifying model-based instrumental norms: model evaluation, argumentative context and instrumental transmission. A notable consequence of this instrumentalist view is that the normative aim of social epistemologists cannot be successfully achieved using exclusively philosophical or exclusively modelling methodologies. Instead, the success of this aim depends on a wide collaboration of epistemologists, modellers, moral philosophers, policy-makers and the like to ensure that reasons for particular epistemic outcomes produce undefeated reasons for socio-epistemic systems.

Keywords— *social epistemology, agent-based models, formal models, normative models, instrumental normativity, socio-epistemic systems, systems-oriented epistemology, network models, landscape models, bandit models*

Contents

1 Introduction

Agent-based models (ABMs) are models that simulate individuals behaviors with the aim of analysing, evaluating and understanding relationships between micro-level behaviors and macro-level phenomena of complex systems. Computer simulation plays a central role in these models: closed-form solutions are usually unavailable and human computing capacities typically fall short of the magnitude of calculations involved in these models [58, 83, 71]. ABMs have found applications in fields as diverse as economics, epidemiology, history, philosophy and the social sciences [46, 16, 96]. Relatively recently, social epistemologists have leveraged the agent-based approach to construct social organisation of scientific research (SOSR) models that aim to understand various relationships between individual epistemic behaviors in socio-epistemic communities and epistemic properties of those communities as a whole. The introduction of ABMs into the epistemologist's toolkit allows for the study of complex socio-epistemic systems, something that was infeasible with earlier philosophical methodologies such as analytical modelling, a priori reasoning and thought experiments [69, 85, 84, 15, 83, 75, 65, 27, 106, 71]. SOSR models form part of what Goldman [27] calls *systems-oriented epistemology*: the branch of social epistemology that examines socio-epistemic systems in terms of their epistemic outcomes. Social epistemologists working within the agent-based approach are interested in, for example, social dynamics of opinion formation [38]; effects of network structure on the speed and accuracy of belief formation of groups [106]; relationships between cognitive diversity and the performance of scientific community [100]; or relationships between trust and belief polarisation in the scientific community [66].

Two central aims can be identified in the agent-based approach of social epistemologists. On the one hand, there is an *epistemic* aim: epistemologists aim to explain, predict and understand the complex dynamics of epistemic communities. On the other hand, a *normative* aim may be identified: epistemologists are interested in using SOSR models in order to provide normative guidance on how to redesign, restructure, or reorganise real-world socio-epistemic systems. In more specific terms, epistemologists pursuing this aim strive to use findings from SOSR models to draw normative conclusions about the organisation of aspects of real-world socio-epistemic communities [15, 92, 19, 35, 31, 72, 9, 21, 70]. Kitcher [45] has given an incisive formulation of the normative aim:

> How do we best design social institutions for the advancement of learning? The philosophers have ignored the social structure of science. The point, however, is to change it [45].

Kitcher's quote suggests that social epistemologists should pay attention to the social structure of science with the aim of changing it to advance epistemic benefits. The normative aim of the agent-based approach is not just an abstract aim of the agent-based approach, but is in fact reflected in the kinds of conclusions that researchers in the field draw from their models. Kummerfeld and Zollman [53] conclude from their ϵ-greedy bandit model that "[i]n some situations, actively encouraging unpopular, risky science would improve scientific outcomes" [53]. To give another example, O'Connor and Weatherall [66] hold that their scientific polarisation models "also suggest a few interventions if, indeed, mistrust of those with different opinions is helping to drive polarization in the chronic Lyme case" [66].

In spite of the interest of researchers and policy-makers in drawing normative conclusions from SOSR models, the normative aim does not come for free: the agent-based approach has been criticised for oversimplifying the true dynamics of systems, a sensitivity to initial conditions, modelling assumptions and decisions, as well as a difficulty in the evaluation process of models [60, 80, 21, 12, 8, 16]. What is more, critics have pointed out that the methodological practices of social epistemologists subvert the connection of SOSR models to real-world socio-epistemic communities and hence the potential for any substantial normative guidance on the organisation of those communities [59, 9, 92]. In response to these worries, we pose the following research question:

> **(RQ)** How can findings from SOSR models normatively inform the organisation of real-world socio-epistemic systems?

In this thesis, we formulate an answer to this question along the lines of instrumental normativity, in which reasons for certain epistemic outcomes as ends produce reasons for socio-epistemic systems as means for achieving those ends. This kind of normativity dovetails well with the kind of findings that SOSR models deliver, which typically describe relationships between micro-level behaviors of epistemic agents and epistemic phenomena emerging on the group level. For the sake of convenience, we refer to instrumental norms based on SOSR models as *model-based norms* in what follows. In answering the research question, we identify three factors on which the justification of model-based instrumental norms depends: model evaluation, argumentative context and instrumental transmission.

Using ABMs to draw up advice for policy design is not a far-fetched endeavour: in fact, it is an exercise in motion. Pandemic ABMs played a substantial role in informing health policies during the COVID-19 pandemic via the mapping out of contagion patterns and the informing of lockdown policies [24, 16, 95, 102, 103]. Other uses of ABMs are in finances and economics [14, 96, 55]; as well as in climate and environmental sciences [34, 5, 1]. Using SOSR models for scientific policy fits well with the twentieth-century development of governmental science policy, in which governments gained interest in shaping "policy discussion about the organization and the funding of scientific research" [70, p. 1]. In fact, SOSR models have already played a role in shaping science-related policies. Bedessem [9] notes that division of labor models in the agent-based approach have been used for providing concrete policy advice about resource allocation and the cognitive structure of communities [9, p. 2]. Hong and Page [39]'s model, for example, "has been cited in support of expected institutional effects of diversity requirements at UCLA, and has been appealed in a brief before the Supreme Court in support of promoting diversity in the armed forces" [31, p. 10]. On a similar note, modelling work on research fund allocation has played a role in supporting the decision-making of various scientific funds: "Explorer Grants" in The Health Research Council of New Zealand; "Seed Projects" of New Zealand's Science for Technology Innovation; and Volkswagen Foundation's "Experiments!" grants [4, pp. 1–2].

Most of the work in the literature has focused on epistemic questions surrounding SOSR models, with two particular lines of thought about the epistemic value of these models. The first line of thought is empirically-based and claims that social epistemologists who want to go beyond toy models and provide insight into real-world socio-epistemic communities should perform systematic empirical validation of their models [59, 85, 92, 21]. The second line of thought denies the need for empirical validation: abstract models can help explore conceptual relationships and thus provide how-possibly explanations, potential explanations, conjectures and novel perspectives on scientific interaction that provide insight into socio-epistemic communities [60, 84]. Epistemic questions about SOSR models, however, fall short of achieving the normative aim: these findings are to be transformed into instrumental norms in which reasons for epistemic outcomes produce reasons for socio-epistemic systems as means to those ends. Roussos [81] and Beck and Jahn [8] discuss accounts that aim to explain normative guidance from formal models in terms of properties of those models. In contrast to these views, our instrumentalist view takes SOSR models as descriptive models that deliver descriptive inputs for instrumental norms. Although the justification of models requires agent-based methodologies, it is argued that philosophical reflection plays a vital role in justifying model-based instrumental norms.

In the following paragraphs, we lay out the structure of the thesis. An outline of the field of agent-based modelling is given in §2, together with a comparison with earlier modelling approaches. As an illustration of the agent-based paradigm, we end the section with Schelling's segregation model. The inception of agent-based modelling in social epistemology is discussed in §2.2 and is emphasised to be systems-oriented approaches within social epistemology. In turn, §2.2.1-2.2.3 provide three instances of SOSR models in order to illustrate the nature and diversity of these models as well as to provide reference material for later chapters. Closing off, §2.3 introduces terminology from the modelling literature for later on and §2.4 discusses the distinction between descriptive and normative models.

In §3, we explain the concept of instrumental normativity, discuss the place of counterfactual dependencies in instrumental norms as well as kinds of normativity failure that occur for instrumental norms. Following this discussion, we first give an overview of Beck and Jahn's concept of normative guidance and discuss three promising accounts of model normativity found in the literature §3.2-3.3. Subsequently, §3.4 argues that the discussed account does not apply to SOSR models in the kind of normativity ascribed to models. Instead, SOSR models ought to be seen as descriptive models, yielding descriptive claims that can be used as input to instrumental norms. This view faces some criticisms from the literature, which we outline in §3.4.2. In the light of these challenges, §6.2 sketches an account of normativity from SOSR models along the lines of instrumental norms.

The role of model evaluation for justifying instrumental norms is discussed in §4. First, §4.1 argues in favor of the importance of model evaluation, building on the criticisms outlined earlier in §3.4.2. Following this, §4.2 discusses the commonplace trade-off between faithfulness and idealising assumptions in the modelling endeavour. In addition, we discuss different kinds of idealisations as well as the question under which conditions idealising assumptions are permissible. In §4.3, we discuss verification, the procedure of evaluating if a model correctly implements it prior conceptual design. Following this,

§4.4 goes over the various kinds of robustness analysis, which each aim to evaluate the stability of a model's results under changes in the model's representation. Third in place, we discuss empirical validation in §4.5, which effectively aims to evaluate the link between the model and the model's target with the help of empirical data and theories. Concluding the chapter, §4.6 argues that model evaluation alone is not enough for achieving the normative aim of SOSR models.

The role of argumentative context for justifying instrumental norms is outlined in §5. First, Aydinonat, Reijula, and Ylikoski's argumentative landscape is outlined, which provides a good starting point for the further discussion. In the next three chapters, we identify an essential as well as a helpful role of argumentative context for justifying instrumental norms. First, §5.2 argues that the justification of instrumental norms depends on an argumentative link connecting broader philosophical concepts to specific conceptualisations that are informally assumed in SOSR models. Next, §5.3 identifies a role of argumentative context for evaluating if a given model is a candidate for justification and, if so, what aspects of the model require veridical justification. In §5.4, we single out an important role of context for robustness evaluation, the discovery of stable results as well as that of so-called difference-makers, which play an important role in constructing instrumental norms. The chapter ends with two criticisms of the argumentative landscape view. First, §5.5 discusses a criticism according to which this view relativises the value of models to their argumentative context. Second, §5.6 examines the view of justification encroachment, which contends that the epistemic standards of justification depend on non-veridical, contextual factors.

At the very end, §6 specifies the role of instrumental transmission in the justification of instrumental norms. First off, §6.1 outlines the function of instrumental transmission from the perspective of the normative aim of SOSR models. This is followed up with an application of normative failures as earlier discussed to the specific context of SOSR models in §3.1.4. In §6.3, we discuss the topic of transmission principles, which aim to stipulate the ways in which reasons for ends transmit to reasons for means. Building on this overview, §6.4 provides a characterisation of the kind of instrumental transmission that we should expect from SOSR models. This leads to the central claim of the chapter in §6.5, namely, that transmission of reasons occurs as long as the descriptive input of the instrumental norm is true, with the strength of reasons depending on the reason for outcomes as well as the system's effectiveness. Although epistemic outcomes thus always produce reasons for socio-epistemic systems facilitating those outcomes, §6.6 identifies three common kinds of defeaters that can undermine those reasons: excessive cost claims, individualist epistemic norms, and moral defeaters.

2 Agent-Based Models

This chapter discusses the paradigm of agent-based modelling, the introduction of ABMs into social epistemology as well as some preliminary concepts. In §2.1, we introduce ABMs and points out some central features of these models, ending with Schelling's segregation model as a clear example of an ABM. Subsequently, §2.2 provides an overview of the introduction of ABMs into social epistemology in the form of SOSR models, emphasising the continuity of the agent-based approach with earlier developments in so-called systems-oriented epistemology. In §2.2.1-2.2.3, three examples of SOSR models are provided, illustrating the nature and diversity of these models and the normative conclusions drawn from these models. In addition, these examples serve as running examples throughout the rest of the thesis. Key terminology from the modelling literature is explained in §2.3, followed by an overview of the distinction between descriptive and normative models in §2.4.

2.1 What are Agent-Based Models?

2.1.1 Agent-Based Models

Agent-based models (ABMs) are computer models with applications ranging throughout disciplines as diverse as economics, epidemiology and social epistemology. ABMs are used for a wide range of purposes, but two overarching purposes can be identified in the background of much of the work. The first purpose for which researchers apply ABMs is to discover or understand complex statistical, causal or counterfactual relationships between local, micro-level interactions amongst agents and environment and macro-level outcomes emerging at the level of the system as a whole. The second purpose of agent-based modelling is to use ABMs to inform choices about the real world. Here, one can think of using ABMs for normative decisions as for instance in policy-making [64, p. 2]. A clear example is the use of ABMs during the COVID-19 pandemic. The Imperial College Model, for example, leveraged ABMs to make forecasts and projections about infection dynamics, which was in turn used to inform policies in the United Kingdom and the

United States [95, 16, 102]. In broad terms, ABMs model complex systems that are composed of *agents*: roughly, entities that act and interact in some given external environment. Schematically, ABMs can be divided into the following building blocks:

(*i*) a set of agents with attributes and behaviors,

(*ii*) a set of agent relations and methods of interaction understood in terms of a topology, and

(*iii*) an environment in which the agents are embedded [58, p. 152].

Although individual agent behaviors tend to follow simple rules, the collective behaviors emerging from interacting agents can be complex and, indeed, unexpected. In other words, properties and actions of individuals can produce large-scale phenomena that those individuals themselves did not necessarily intend. In the literature, the term *emergence* is sometimes used to describe how seemingly simple micro-level interactions in ABMs result in unintended, unforeseen macro-level consequences [58, 83, 69].

Notably, ABMs form a departure from earlier analytical approaches that use equation-based models (EBMs): sets of equations representing macro-level relationships between variables of interest. EBMs typically allow for an analytic solution, are efficient to run and relatively simple to understand. This allows the modeller to control, manipulate and better understand the equations involved. An example of an equation-based model is the so-called SIR model, which models a population of susceptible individuals (S), infected individuals (I) and recovered individuals (R) using the following three differential equations:

$$\frac{dS}{dt} = \frac{-\beta SI}{N}$$
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I,$$

where β is the rate of transmission per contact, γ the recovery rate and N the population size [42, pp. 3–4]. Without going into the details, the SIR model illustrates the core of equation-based models: the model's representation lies at the level of macro-level properties, namely, the rates of change in susceptibility to disease, infectedness and recovery in the population. In contrast to this focus on macro-level properties, agent-based models start at the level of micro-level dynamics, which specify the behavior of individual agents with respect to the environment and other agents. In turn, these micro-level dynamics are observed to bring about macro-level outcomes which can be linked back to the initial micro-level dynamics. Here, a typical agent-based approach to modelling infectious diseases would model the micro-level behaviors of individuals and link these behaviors to macro-scale infection patterns. Since a population is modelled as a collection of agents that have their own properties and decision-making rules, the agent-based paradigm allows for modelling heterogeneous populations. In this, ABMs contrast with EBMs, which typically assume homogeneous populations underlying the macro-level relationships [83, 36, 42, 71]. A clear example of this is given in the SIR model, where transmission and recovery rates are assumed to be the same across all individuals in the population. As a caveat, ABMs are highly complex due to stochastic components and non-linear behavior, which is often difficult to solve analytically and requires the help of computer simulations [71, p. 54].

Condensing the agent-based approach, Epstein [20] lists five characteristics that are central to agent-based modelling:

(*i*) *Heterogeneity*: agent can differ in their individual properties as well as their relational properties to other agents and environment.

(*ii*) *Autonomy*: agents make their own decision to achieve and further goals.

(*iii*) *Explicit space*: agents are situated within some environment.

(*iv*) *Local interactions*: agents interact with other agents close to them.

(*v*) *Bounded rationality*: agents act on the basis of simple rules and locally available information.

Taken together, these aspects of agent-based modelling allow for the modelling of complex dynamics that were infeasible with the tools of the earlier equation-based approaches. In addition, the bottom-up perspective of the agent-based approach helps to provide insights into the process that gives rise to macro-level emergent phenomena. As an example, the agent-based approach allows for studying ways in which individual actions based on limited information may contribute to the spread of contagious disease [20, 69]. ABMs model systems on different levels of abstractions. Squazzoni, Boero, et al. [88] introduced an insightful classification of ABMs according to the target systems represented, generalised to the context of social epistemology by Šešelja [85]:

(*i*) *Case-based models*: aims to represent concrete cases from scientific practice, taking into account appropriate empirical information about those cases.

(*ii*) *Typifications*: aims to represent a "*class* of empirical phenomena" by considering certain properties of that class "while abstracting away from particularities of each individual phenomenon".

(*iii*) *Theoretical abstractions*: highly simplified and idealised models that aim at a simple, general representation of phenomena, often with an exploratory function, e.g., for "testing new ideas, extending existing frameworks" [85, pp. 3–4]

Although one might assume that the move from more complex to more simple models corresponds to a move from more specific to more general models, Šešelja [85] and Edmonds and Moss [18] warn against this assumption: "simplifying a model won't necessarily make it more general". Simplifying a model, for example, could result in ignoring features highly relevant to the modelled system, making the model less general [88, 85, 83].

Closing off this section, it should be noted that the practical value of agent-based models has been criticised by researchers and philosophers alike. Douven [16] describes three commonly raised criticisms of the agent-based approach:

(*i*) The tendency to oversimplify agent behavior, decision-making processes and environmental processes may result in a failure to capture the true dynamics of a system.

(*ii*) The sensitivity to initial conditions, modelling assumptions and decisions make it difficult to draw general conclusions from models.

(*iii*) The difficulty of evaluation due to model complexity as well as the shortage or inaccessibility to data.

In one sense, all of the raised criticisms applies to models in the generic sense: every model involves a risk of misrepresentation, non-generalisability as well as complexity or the unavailability of data. In fact, ABMs even promise to overcome some of the problems of earlier analytical approaches, most notably by weakening assumptions of unbounded rationality and homogeneity. Taken in another sense, the risks noted may come up in different or even novel forms for the agent-based approach and should, at any rate, be taken as serious risks. Although the points raised are not to be taken as fundamental criticisms of the agent-based approach, we do agree that the points draw attention towards the pitfalls of agent-based modelling and, therefore, call for diligence [16, p. 34].

2.1.2 Schelling's Segregation Model

Historically, an illustrative example of an ABM has been Schelling [82]'s residential segregation model, introduced into sociology in the 1970s.[2] Consider a grid of cells, a tolerance parameter $p \in [0, 1]$ and two same-sized groups of agents that each occupy exactly one cell on the grid. In the initial step, agents are randomly distributed over the grid. On each successive round, every agent checks their directly neighboring cells and evaluates if the fraction of occupied cells containing individuals of one's own group is at least p. If so, then the agent stays in their respective cell. If not, then the agent tries to find the nearest uninhabited cell that meets this demand [82, pp. 154–155]. The simulation run ends if no move can increase the number of satisfied individuals [87, p. 343]. Even in cases where the tolerance parameter is low, Schelling's model surprisingly results in highly segregated final states [82, 67]. Figure 1 illustrates a successive run of Schelling's model on a 51×51 grid with $p = 0.33$, successively moving towards a segregation pattern. Next to providing a simple example of an ABM, Schelling's model illustrates an important point of ABMs. Although the micro-level actions of individuals did not aim at segregation, this phenomenon did result from their collaborative efforts. Generally speaking, Schelling's model illustrates what is called

[2] Although Schelling [82] introduced two models in their original paper "Dynamic Models of Segregation", we discuss the second, two-dimensional model that is often simply referred to as Schelling's residential segregation model.

the concept of emergence as described in §2.1. Running ABMs provides fertile ground for more understanding, novel perspectives, conjectures as well as how-possibly explanations. While how-actually explanations aim to pinpoint the actual causes of a given phenomenon, how-possibly explanations instead provide conditions that may give rise to that phenomenon. Schelling's model belongs to the category of how-possibly explanations: it provides conditions under which segregation may occur but does not show that these conditions provide a causal explanation of actual practices of segregation [82, 67, 32].

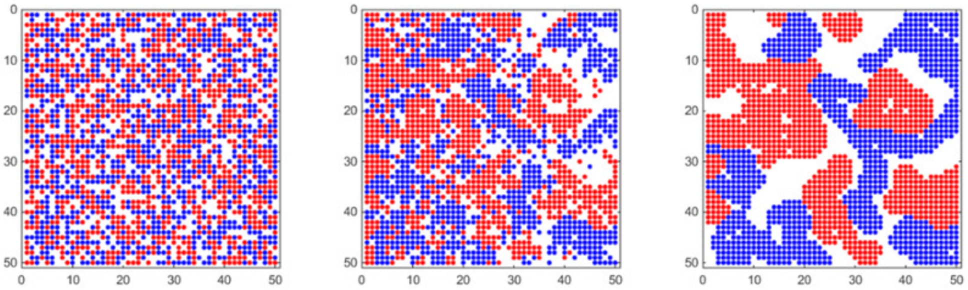

Figure 1: Run of Schelling's model on a 50×50 grid with $p = 0.5$ at timesteps $t = 0$, $t = 4$ and $t = 30$, after Rafał Kowalski.

2.2 The Place of Agent-Based Models in Social Epistemology

Social epistemology is a branch of epistemology that breaks with traditional epistemology by shifting the focus away from the epistemic individual towards epistemic outcomes of social practices, norms and systems. Goldman [27] carves the field up into three different lines of research:

(*i*) the study of *individual doxastic agents* with social evidence,

(*ii*) the study of *collective doxastic agents*, and

(*iii*) the study of *socio-epistemic systems*.

Prominent topics of research within social epistemology are testimony, peer disagreement, epistemic relativism and judgement aggregation; each of these topics can be categorised under one of Goldman's groupings [28, 65]. An important motivation for social epistemology's focus on groups is what Mayo-Wilson, Zollman, and Danks [61] call the *independence thesis*, which claims that there is a logical independence between what is epistemically rational for individuals and what is epistemically rational for groups [61, pp. 653–654]. On the one hand, rational individuals can form irrational groups. On the other hand, rational groups may consists of individuals that nonetheless act in irrational ways. As an example, it can be epistemically rational on the group-level if members of that group cling to a theory even in the light of strong evidence against that theory [61, pp. 653–655]. Another example is List and Pettit [57]'s impossibility theorem about judgment aggregation, which hold that there exists no aggregation procedure turning rational sets of individual judgments into a collectively rational set of judgements given plausible constraints on the aggregation procedure [57]. Brushing over the details, List and Pettit's result illustrates that individually rational judgments need not bring about collectively rational judgments. Since epistemic prescriptions for individuals may not generalise to groups, the role of group organisation in guiding reliable, correct inference and belief-forming on the group-level is a separate subject of study for social epistemologists over and above the individualist perspective [61, 75, 106, 15].

In what follows, we focus on the third variety of social epistemology, which Goldman [27] labels as the *systems-oriented variety*. This variety studies social systems in terms of the epistemic outcomes of these systems on their members. Systems of interests are typically formal institutions with specified aims, rules and procedures: the scientific enterprise, the education system, journalism and so forth. Systems-oriented epistemologists work within a broadly consequentialist framework: particular social systems are evaluated in terms of the extent to which they promote or impede valued epistemic outcomes. In this endeavour, epistemologists ascribe more epistemic value to systems that generate better epistemic outcomes. This consequentialist approach is to be contrasted with non-consequentialist approaches in epistemology that draw on formal theories or models to work out norms that apply regardless of their consequences [27, p. 14]. Social epistemologists in the systems-oriented tradition study social systems with at least two aims in mind. The first aim is descriptive and aims to verify if the mode of operation of these systems, in fact, furthers the epistemic ends that the groups purport to pursue. Zollman [106], for example, has suggested that furthering accuracy of beliefs in some scientific contexts may require scientists to be "uninformed about experimental results in their field". This can be taken to suggest that organisational structures in which scientists are informed about all results may sometimes not further the epistemic ends of optimising accuracy of beliefs [106, p. 586]. The second aim is normative: epistemologists working within this aim attempt to make conjectures, finding potential explanations and drawing conclusions about factors that contribute to some desired epistemic outcome in order to redesign, restructure, or reorganise socio-epistemic communities so "that [these] communities can thrive in epistemically relevant ways" [15, p. 459]. Here, Zollman's result may be taken as a suggestion for redesigning how communication is organised within the scientific enterprise.

In recent years, the agent-based approach has become a well-established tool in social epistemology for studying descriptive and normative questions about socio-epistemic systems [84, 83, 65, 15]. It is insightful to reflect on the place of ABMs in the field of social epistemology. On the one hand, ABMs as computer models form an obvious discontinuity with established, mainstream philosophical models such as a priori reasoning, arguments and thought experiments [46, pp. 9–10]. On the other hand, ABMs form a continuity with formal tools that were already part of the epistemologist's toolbox. Philosophers have long studied epistemological issues with the help of logic, decision theory and Bayesian probability. From this perspective, ABMs expanded these formal tools and brought novel possibilities for studying more complex phenomena [56, 60]. An illustrative example of the continuity between ABMs and earlier formal tools are division of labor models. In their work, Kitcher [45] introduced an analytical model with the aim of examining optimal allocations of research resources between competing lines of research.[3] The model suggested that individual epistemic rationality is neither necessary nor sufficient for collective epistemic rationality. On the one hand, individual rationality is not necessary for the following reason: non-epistemic incentives such as self-interested credit-seeking can open up unexplored research directions with a low chance of success but high epistemic returns, which can result in a collectively optimal resource allocation. On the other hand, it is not sufficient: it is rational for individuals to focus on research pathways with a high chance of success, but this can prematurely rule out research pathways with low chance of success but high epistemic returns [45, 83]. Muldoon and Weisberg [63] adopted an agent-based approach to drop Kitcher's two controversial assumptions from Kitcher's analytical model, which allowed

[3] Kitcher's model falls under economics-inspired approaches focused on studying relationships and tensions between individual and collective rationality through the lens of the *generous invisible hand*, according to which interaction between individuals produce collective-level consequences, without individuals necessarily aiming at those consequences.

Muldoon and Weisberg to study optimality of divisions of labor under less stringent assumptions. Although we further discuss Weisberg and Muldoon's model in §2.2.2, we note that their work showed that Kitcher's results were highly dependent on the assumptions in Kitcher's analytical model [63, 83]. Next to the continuity with earlier approaches, ABMs fit well with the natural interests of social epistemologists, particularly those working within the systems-oriented approach: their interest is directed towards understanding epistemic phenomena of socio-epistemic systems [92, p. 315]. A particular interest of social epistemologists is in the social practice of modern science, as Douven [15] aptly expresses in the following quote:

> [I]t is somewhat of a commonplace that modern science can only be understood as an essentially social enterprise, involving groups of interacting agents who must be willing to circulate results and must be open to other manners of collaboration [...] [15, p. 459].

Needless to say, ABMs are a natural candidate for studying questions about complex socio-epistemic systems and their dynamics [15, p. 457]. As Mayo-Wilson and Zollman [60] further argues, properties of socio-epistemic systems discredit more traditional traditional methodologies of a priori reasoning, arguments and thought experiment as good practice. In particular, traditional philosophical methodologies are not equipped to deal with the complexity, number of variables and stochastic nature inherent to socio-epistemic systems. On the contrary, agent-based modelling specifically aims to deal with complexity, dimensionality and stochasticity and thus forms a more reliable paradigm for effectively studying these systems [60, p. 3658].

2.2.1 Zollman's Epistemic Network Model

Zollman [106]'s *epistemic network model* (ENM) is focused on the role of social networks in science. Given the differences in what is epistemically rational for individuals and groups, Zollman argues, it is a relevant question to ask what kind of social network is best for achieving outcomes that are epistemically rational on the group-level. Zollman's model suggests that there exists a trade-off between the speed at which groups form beliefs and how accurate those beliefs are. As we will discuss, this result is not robust across parameter space but occurs when learning is more difficult. To introduce the model, Zollman asks us to imagine a group of scientists who have found a new treatment to some medical condition. Although the researchers are well aware of the success rate of the old treatment, there is uncertainty about the effectiveness of the new treatment. Through experiments and communication, scientists collect data to form and update their beliefs about the new method's effectiveness. Their aim is to find the method that is most effective [97, 83, 106, 105, 80]. Zollman models this situation using what is called a *n-armed bandit problem*. Analogous to a slot machine, an agent in this problem can pull one of n levers and observes rewards from that pull. Through repeated plays, the aim is to maximise by playing the levers with the highest expected reward. Since the optimal action is unknown to the agent, they have to explore the available options and ultimately exploit the best available option [106, 91].

In Zollman's model, experiments are seen as analogous to pulls or actions and experimental results as analogous to rewards or payoffs. In addition, agents communicate their findings to some other agents. More formally, k agents are situated on an undirected graph interpreted as a communication structure as shown in Figure 2. In the initial step of a simulation, each agent is assigned a prior credence, interpreted as the belief that actions B is better than action A. On each further round, agents can take one of two actions: A or B, with respective success probabilities $p_A = 0.5$ and $p_B = 0.5 + \epsilon$ for some fixed $\epsilon \in [0, 0.5]$. In deciding an action, agents use the following decision rule: perform action A if the credence that B is better than A is lower than 0.5, otherwise perform action B. On performing an action, agents receive a payoff from a binomial distribution with parameters p and n, with p the success probability of the agent's corresponding action and n a parameter determining the number of experiments agents perform on a round. On observing their own payoff as well as the payoffs from neighbors, agents perform Bayesian conditionalisation to update their credences. The model simulation ends if either of two conditions is met: false consensus, in which everyone chooses action A so that evidence for B cannot be obtained anymore, or true consensus, where all agents have a sufficiently high credence that action B is better than action A [106, 97, 80].

The left-hand plot in Figure 3 gives an interesting and surprising result: faster convergence on beliefs can be detrimental for belief accuracy whereas slower convergence can produce more accurate beliefs. This has been dubbed the *Zollman effect*: more communication can be detrimental to knowledge discovery. Zollman explains this effect as follows:

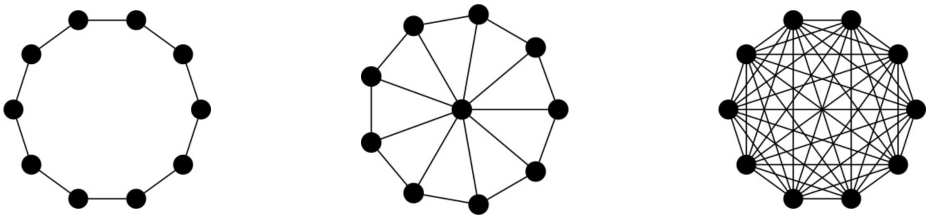

Figure 2: Cycle, wheel and complete graphs, after Šešelja [83]. In the cycle, every individual is connected to two other individuals; the wheel is the cycle plus an individual that is connected to everyone; in the complete graph, everyone is connected to everyone.

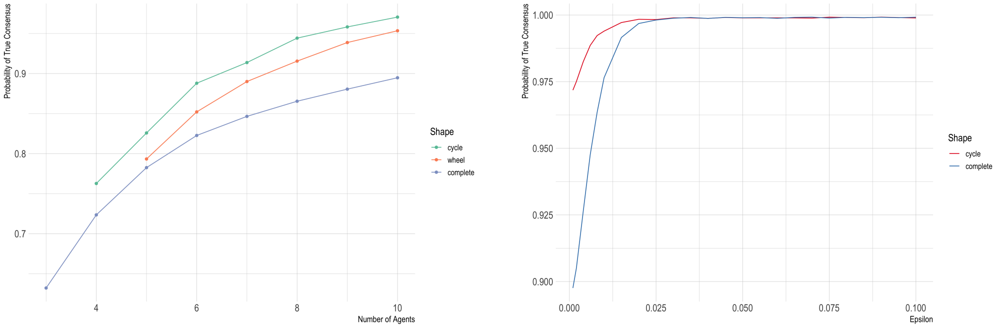

Figure 3: Experimental results from Zollman [106] and Rosenstock, Bruner, and O'Connor [80] respectively, after Weisberg [97]. In both plots, the vertical axis gives the probability of true consensus defined as the fraction of 10,000 simulations ending in true consensus; the lines corresponds to graph types; the horizontal axis in the left plot gives values of k whereas the horizontal axis in the right plot gives values of ϵ.

> [U]nconnected networks are more robust to the occasional string of bad results than the connected network because those strings are contained in a small region rather than spread to everyone in the network. This allows the small networks to maintain some diversity in behaviors that can result in the better action ultimately winning out if more accurate information is forthcoming [106, p. 583].

Hence, more connected networks are prone to a kind of epistemic pollution: because everyone is connected to everyone, spurious results will spread more quickly throughout the whole network. Contrastingly, less connected networks are less prone to epistemic pollution: since individuals are isolated, a diversity of beliefs emerges, which reduces the risk of converging to the wrong belief. According to Zollman, there is a robust trade-off between speed and accuracy: more connected networks will converge more rapidly on the hypothesis but are more likely to converge on the wrong hypothesis and less connected networks converge less rapidly but are less likely to converge on the wrong hypothesis [106, 80, 83, 97, 69]. Since epistemic pollution can result in the community settling on the wrong belief early on, Zollman underlines the benefit of what is called *transient diversity*: the inquiry process in which communities keep exploring the available theories sufficiently long so as to ensure that the best of available theories is not prematurely abandoned [107, 83].

Rosenstock, Bruner, and O'Connor [80] re-evaluated Zollman's model and disproved the claim that the trade-off between speed and accuracy is robust across parameter space. In particular, Rosenstock et al. found that the Zollman effect mostly occurs in regions of parameter space that make the learning problem especially difficult: when the success rates between hypotheses are similar, when the number of agents is small, or when the amount of collected data is small [80, p. 248]. To illustrate their explanation, Figure 3 shows that as the difference between success probabilities ϵ grows, performance between the cycle and the complete graph becomes increasingly similar. Effectively, higher success rates make it easier for agents to identify B as the correct action from received payoffs, with the effect of communication structure becoming smaller and smaller as success rates increase. In spite of their robustness criticism, Rosenstock et al. agree with Zollman that transient diversity is important for optimal epistemic outcomes. Transient diversity, they argue "is necessary to good scientific practice" and "holds for all the models discussed" in their article [80, p. 251]. Interestingly, Rosenstock et al. further disagree with normative claims that Zollman draws from the model. Zollman [106] namely suggests that in some cases, information flow should be decreased between agents so as to improve epistemic diversity and optimise the fruits of collective inquiry [106, pp. 585–5 86]. This verdict strikes Rosenstock et al. as unintuitive: in cases where "good data are hard to come by [...] there are better solutions [...] to the problem". In fact, Rosenstock et al. argue that limiting information can epistemically be harmful: scientists should consider all the available data in judging which theory is the best theory. To avoid epistemic pollution, Rosenstock et al. suggest some options that could increase the chances of getting it right: the inclusion of "more stubborn or exploratory agents, who keep investigating both theories, even if one seems less promising, will not preemptively settle on a poor theory"; as well as the establishment of "[s]tandards for the amount of data necessary to confirm or deny a theory" [80, p. 251].

2.2.2 Weisberg and Muldoon's Epistemic Landscape Model

Weisberg and Muldoon [100] introduced their *epistemic landscape model* (ELM) to study how the selection of research approaches affect distributions of cognitive labor [69, 100]. Weisberg and Muldoon's model departed from earlier models inspired by Kitcher, which make the following two assumptions:

(*i*) *The distribution assumption*: each agent knows the distribution of cognitive labor before choosing to work on a project.

(*ii*) *The success function assumption*: each project is assigned an objective success probability known by all agents [45, 63, 69].

By taking an agent-based approach, Weisberg and Muldoon can drop these two assumptions: agents have varying access to information and follow their own behavioral rules [71, pp. 51–52]. Two central conclusions that Weisberg and Muldoon draw from their model is that, first, researchers work best if they take into account social information and, secondly, a mix of conservative and risk-taking research approaches is optimal [100, pp. 250–251].

Weisberg and Muldoon's model builds on two assumptions: first, future research builds on earlier research and, second, progress in a given research group can affect the research trajectories of other groups. Conceptually,

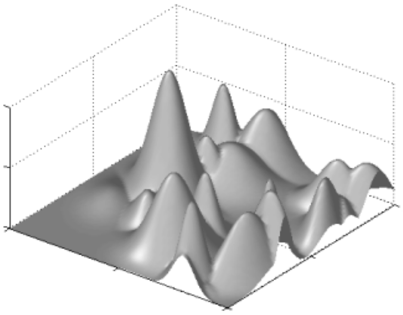

Figure 4: Example of a three-dimensional epistemic landscape model, after O'Connor, Goldberg, and Goldman [65].

Weisberg and Muldoon interpret a particular topic of research as being explored by researchers who built on each other's work and possibly influence the approaches taken by others. In this regard, researchers can adopt different approaches with some approaches more epistemically significant than other approaches. ELMs formalise this conceptual model. A *research topic* is formalised as an *epistemic landscape*, which is effectively a n-dimensional space. In this space, $n-1$ dimensions encode the *research approach*; the last dimension is reserved for the *epistemic significance* of that approach. Higher values correspond to higher epistemic significance; agents use different search strategies to explore the landscape with the ultimate aim of reaching points with high epistemic significance in the landscape's curvature [65, 100]. To visualise the model, a three-dimensional landscape example is provided in Figure 4. In this example, a research approach is a two-dimensional coordinate on the surface, with the corresponding height designating the epistemic significance of the given approach. Observing the ruggedness of this particular landscape, one can imagine that optimally exploring the landscape can be difficult: depending on the way in which one explores the landscape, one might settle on a point with less than peak epistemic significance.

To explore the landscape, agents must have some kind of search strategy that determines their exploration behaviors. Weisberg and Muldoon [100] examine three kinds of agents with their own strategies for exploring the landscape:

(*i*) *Controls*: move towards points with higher epistemic significance, whilst ignoring the exploration done by others.

(*ii*) *Followers*: move towards points with higher epistemic significance that

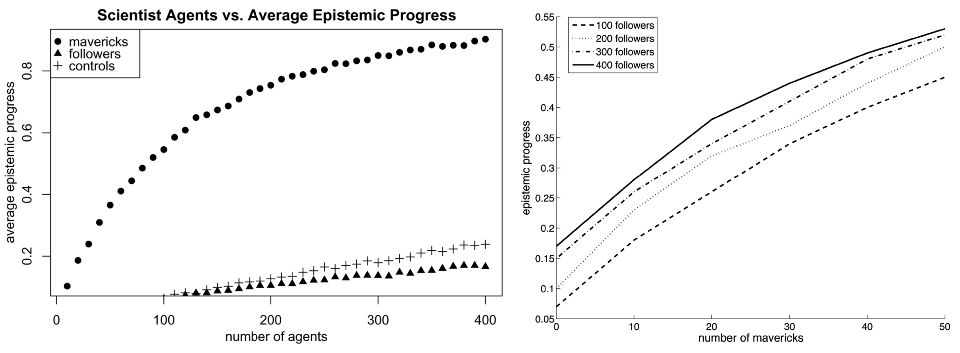

Figure 5: Experimental results from Weisberg and Muldoon, after Weisberg and Muldoon [100]. In the left-hand plot, the horizontal axis gives the number of agents in a population; the vertical axis gives the average epistemic progress; the lines correspond to the different types of agents. In the right-hand plot, the number of mavericks in a population is displayed on the horizontal axis; the epistemic progress on the vertical axis; the lines correspond to the number of followers in the population.

have already been explored by others.

(*iii*) *Mavericks*: move towards points with higher epistemic significance, whilst avoiding points that have already been explored by others.

In colloquial terms, controls are effectively lone scientists that ignore work from others. In contrast, followers and mavericks both use social information in their own way: whilst followers follow the track of research already pursued, mavericks observe the research paths undertaken and aim to innovative on what has already been done [100, 83, 32, 29].

Weisberg and Muldoon investigate the relationship between the selection of research approaches and optimal distributions of cognitive labor by evaluating the effects of homogeneous and mixed populations of controls, followers and mavericks on epistemic progress [83, 100]. In the experiments, Weisberg and Muldoon interpret epistemic progress as the "the percentage of patches with significance greater than zero that have been visited by the community of scientists" [100, p. 232].[4] Figure 5 shows results from Weisberg and Muldoon's experiments. In the left-hand plot, it can be observed that in homogeneous populations, mavericks substantially outperform both

[4] Aydinonat, Reijula, and Ylikoski [6] have questioned the construct of epistemic progress that Weisberg and Muldoon employ [6, p. 381]. One clear problem is that Weisberg and Muldoon's notion of epistemic progress is insensitive to the sum of epistemic significance of visited patches. This may be relevant if we want to compare two groups that have explored the same percentage of patches with non-zero significance.

followers as well as controls. The right-hand plot shows mixed populations of followers and mavericks. Across the spectrum, an increase in the number of mavericks is seen to improve epistemic progress. According to Weisberg and Muldoon, the increase in performance is due to both a direct and indirect influence of mavericks. First of all, mavericks themselves are already better at finding high significance approaches. Secondly, mavericks can help followers get unstuck from unfruitful approaches by leading the way [100, pp. 246–247]. It has to be noted that further work from Alexander, Himmelreich, and Thompson [2] has dismissed this second claim as the artifact of an implementation error of Weisberg and Muldoon's model, which we further elaborate on when we discuss verification in §4.3 [2, 93, 32].

Although Weisberg and Muldoon do emphasise that their models are highly idealised, they draw two broader tentative conclusions for the division of cognitive labor. First of all, it is key that researchers do not ignore the paths taken by other researchers. Controls ignore social information and, as a result, receive suboptimal outcomes when compared with followers and mavericks. Secondly, an optimal division of labor requires "a polymorphic population of research strategies": a mix of conservative and risk-taking researchers is conducive to epistemic progress. Although more mavericks are always optimal, Weisberg and Muldoon do emphasise that "[d]ifferent strategies have differential costs" and that "it is more costly to be a maverick than a follower": their anti-conservative attitude to exploring the landscape gives mavericks a larger burden of research [100, pp. 250–251]. Their view is that in real-world socio-epistemic communities, optimal research communities will plausibly have a smaller number of mavericks due to the high costs of adopting the maverick's attitude [69, 100, 65].

2.2.3 Kummerfeld and Zollman's ϵ-Greedy Model

Kummerfeld and Zollman [53] introduce their ϵ-greedy model to examine an assumption about the *scientific state of nature*: namely, the assumption that scientific progress is most efficient if scientists are completely left to their own devices and thus free from outer institutional control. Contradicting this assumption, Kummerfeld and Zollman's model suggests that there is an efficiency in the scientific state of nature that takes the form of a free-rider problem: in the state of nature, scientists are incentivised to pursue the most promising lines of researchers and leave the exploration of novel research avenues to other scientists. As a consequence, Kummerfeld and

Zollman argue, “one cannot simply presume that science is most efficient when institutional control is absent” [53, pp. 1057–1058]. Kummerfeld and Zollman's model has a similar setup to Zollman [106]'s model: agents take actions, receive payoffs from those actions reflecting success and failure of those actions and aim to find the action that provides the highest average payoff [53, pp. 1059–1060]. A difference with Zollman [106]'s model is the use of ϵ-greedy strategies for decision-making. Consider two types of agents: *exploiters* and *explorers*. On the one hand, exploiters take actions with the highest average payoff.[5] On the other hand, explorers do not necessarily choose the highest average payoff action but instead alternate between the available actions obtain a better picture of the payoff of different actions. This points at something called the *exploration/exploitation trade-off*: exploiting may result in suboptimal actions because promising alternatives are ignored, whereas exploring can be suboptimal because the optimal action has already been found, making further exploration unnecessary [53, 91]. An ϵ-greedy strategy is a strategy in which agents take actions according to the following decision rule: given fixed $\epsilon \in [0, 1]$, take the highest average payoff action with probability $1-\epsilon$ and alternate to some other action otherwise. The parameter ϵ is called the *exploration rate* and effectively determines the explorative and exploitative tendencies of agents, with higher values corresponding to more exploration and lower values to less exploration. Kummerfeld and Zollman interpret ϵ as a mix of personal and external circumstances and as “a measure of how prone the scientist is to pursuing risky lines of research” [53, pp. 1061–1062, 1063–1065].

Kummerfeld and Zollman evaluate the efficiency of the state of nature in an experiment. In the experiment, agents pull fifty times from a two-armed bandit, with the first arm always returning a payoff of zero and the second arm returning some value from a normal distribution with mean zero and variance nine. Next, Kummerfeld and Zollman proceed to ask the following question: what value of ϵ is best to adopt? On the group-level, the optimal ϵ is interpreted as the value that maximises the sum of utilities if all scientists would adopt that ϵ. On the individual level, the best ϵ is interpreted as the best value that one could respond with given that the group plays some fixed ϵ'. From this, Kummerfeld and Zollman infer the symmetric Nash equilibria on the individual's best responses: intuitively, these are the exploration rates

[5] In Zollman's model, agents follow an exploitative strategy: their decision rule is to choose the action that they believe to be optimal.

for which the individual and the group play the same strategy and are not incentivised to switch strategies. Although the further details are not crucial to the discussion, Figure 6 shows the symmetric Nash equilibria for each group size, which is unique and defined by the points where the best response value equals the group value for the exploration rate. More details on how the authors inferred the symmetric Nash equilibria can be found in Kummerfeld and Zollman [53]. Figure 7 shows the differences in values of the optimal exploration rate as compared with the group-level and the symmetric Nash equilibria for both the complete graph and the cycle graph. In both cases, it is clear that symmetric Nash equilibria consistently give a lower exploration rate than community optima. In other terms, the best strategy for individuals is to experiment less than is optimal [53, p. 1067]. Hence, in the scientific state of nature, the best strategy is to explore less, resulting in less optimal epistemic outcomes than achievable. Kummerfeld and Zollman recognise this as a free-rider problem: the state of nature incentivises individual scientists to conservatively exploit the most promising research avenues whilst leaving the burden of exploration to other scientists [53, p. 1058]. In the quote below, Kummerfeld and Zollman draw a normative conclusion from their findings:

> This inefficiency suggests that one cannot simply presume that science is most efficient when institutional control is absent. In some situations, actively encouraging unpopular, risky science would improve scientific outcomes [53, p. 1057].

2.3 Frameworks, Models and Targets

In the coming chapters, some terminology from the modelling literature will be helpful. The literature is filled with different conceptualisations of models: models have been viewed of as sets of sentences, mathematical structures, fictions, caricatures and more [22, 92]. Although a full-blown characterisation of SOSR models goes beyond the scope of this thesis, we will loosely think of them as mathematical or computational constructs that aim to represent some aspect of socio-epistemic inquiry [92, p. 315]. On the received view in the philosophy of modelling, models should be distinguished from their *target*: the intended or unintended referent of the model or, alternatively, the system that the model represents. Using the model to draw conclusions about the target, in turn, requires that the model and target are linked such

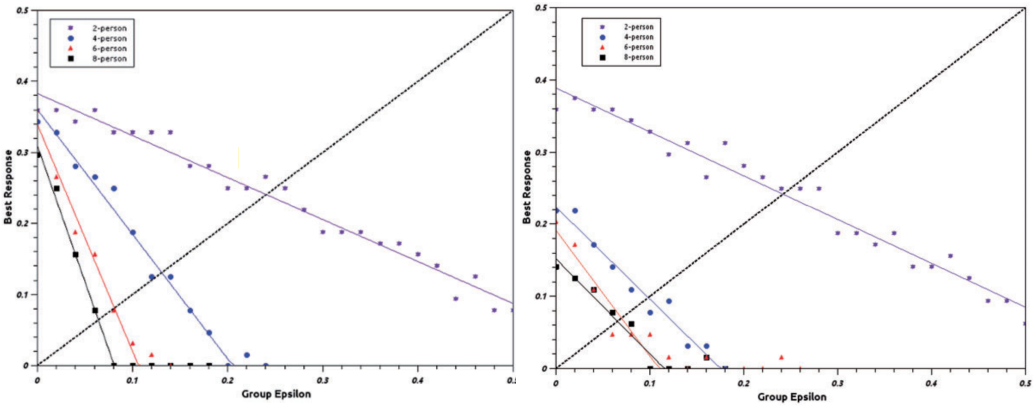

Figure 6: Linear regressions for the complete graph and cycle respectively, after Kummerfeld and Zollman [53]. In both plots, the horizontal axis gives different values of the exploration rates for the group; the vertical axis gives the individual's best response exploration rate; the solid lines are regressions for coupled values of group and best response rates for each group size; and the dotted line gives the symmetric Nash equilibria for each group size.

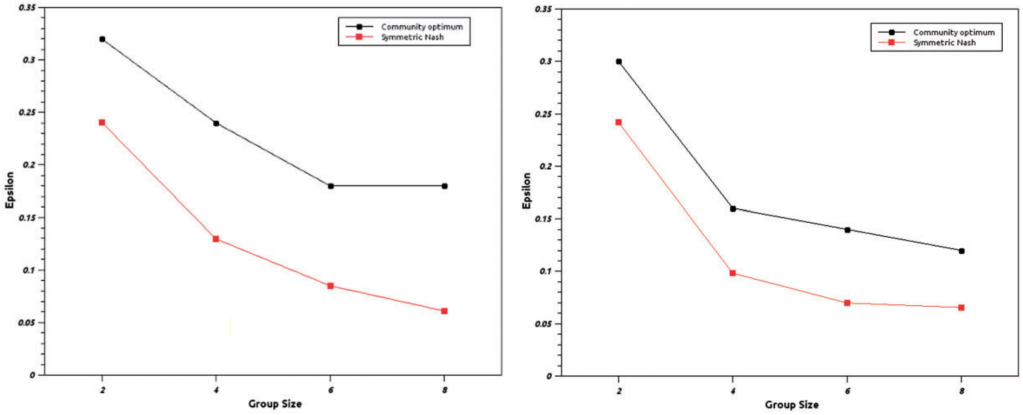

Figure 7: Experimental results for the complete graph and cycle respectively, after Kummerfeld and Zollman [53]. In both plots, the horizontal axis provides the size of the group for groups sizes; the vertical axis gives different values of the exploration rate ϵ; and the lines correspond to community optima and symmetric Nash equilibria.

that inferences made about the model reliably transfer to the target system [22, 59, 85, 21]. Since SOSR models aim to represent some aspect of socio-epistemic inquiry, there is always an intended target, namely, the aspect of socio-epistemic inquiry of which the model is a model. This means that a construct without target is not properly called a *model*, although the status of a model could be received if a target would be fixed.

Targets come in different varieties depending on the level of abstraction. Gräbner [30] provides a fruitful classification that classifies targets from concreteness to abstractness:

(*i*) *Concrete targets*: particular, real-world systems delineated in time and space.

(*ii*) *Generalised targets*: abstract systems that represent features common to a class of systems.

(*iii*) *Hypothetical targets*: non-existing systems, mostly used for providing potential explanations [30, p. 5].

This classification matches up with the one given earlier in §2.1.1: case-based SOSR models pair with concrete targets; typifications with generalised targets; theoretical abstractions with hypothetical targets. The kind of target involved matters for drawing conclusions for real-world epistemic communities: applying conclusions about concrete targets to generalised targets may be a too hasty generalisation whilst attributing properties of hypothetical targets to concrete targets can fail if the hypothetical target differs too much from the concrete target [85, p. 541].

An indirect form of inquiry is central to the modelling endeavour: instead of directly investigating the target system, the modeller investigates the model and makes inferences about the model. In turn, inferences are transferred to the model's target [81, 79, 23, 12]. Before obtaining a model, the modeller requires what Titelbaum [94] calls a *modelling framework*. Effectively, a modelling framework is the model's formal structure without assigning values to the model's parameters. Consider, for example, Zollman's ENM. Here, we can think of the modelling framework as the formal learning structure introduced by Bala and Goyal [7]. In turn, a model is obtained by interpreting the kind of graph, the number of agents k, the number of experiments n and the success probability parameter ϵ. Titelbaum's distinction is pertinent to the model-target link. On the one hand, a model's interpretation allows for infusing the model with information from the real-world target: the modeller can use information from the real world in setting the values of the model's parameters. In addition, the model's predictions can be tested against information obtained from the real-world target. As Titelbaum [94] notes, interpretation is a two-way street from world to models and vice versa: "we need to know not only how features of the world should be captured in features of our models, but also how to read off features of models as conclusions about the world" [94, pp. 4–7].

2.4 Descriptive and Normative Models

In the literature, philosophers distinguish *descriptive models* from *normative models*. On the one hand, *descriptive models* are usually understood as models serving some kind of descriptive aim such as explanation, prediction or understanding. On the other hand, the debate surrounding normative models is more complicated. Although philosophers endorse a number of adjacent yet competing conceptions of what normative models are, there is agreement that the very same construct can be taken as descriptive or normative depending on the purposes to which the model is set [8, 12]. Take, for example, the transitivity of preferences axiom in expected utility theory. On the one hand, one may interpret this axiom as a descriptive claim about people's psychology, which can be used for descriptive purposes such as predicting people's actual preferences. On the other hand, transitivity of preferences may be seen as a prescriptive constraint on reasoning and the organisation of beliefs, which can be used for normative purposes such as deciding which action is preferable from a given set of actions. In a similar vein, the modus ponens rule may be taken either as a descriptive fact or as a normative constraint on reasoning [12, p. 1342].

Titelbaum [94] defines normative models as models that "attempt to fit normative facts" such as prescriptions, evaluations and general facts about normative concepts. A model that fits reasonably well to the normative data may, in turn, issue normative verdicts through prediction [94, p. 3]. Beck and Jahn [8] provides an alternative view of the normative of models, according to which models are normative if their verdicts can be taken as recommendations for action to which agents can appeal [8, p. 131]. In a similar vein, Colyvan sees normative models as models that prescribe "how agents *ought* to act": such models "are taken to prescribe how we ought to reason, organise our beliefs, and so forth" [12, pp. 1338, 1340].

3 Instrumental Normativity

In this chapter, we discuss instrumental normativity, normative guidance, accounts of model normativity and our account of model-based instrumental norms. In §3.1, we provide an outline of instrumental norms as norms that prescribe taking means that further one's ends, based on a counterfactual dependency between means and ends. Moreover, we identify two ways in which reasons for means can fail to be normative. Following this discussion, §3.2-3.3 provides a literature overview, in which we explore how model normativity has been construed as well as three accounts that aim to explain how models can exert normativity. In §3.4, it is argued that these accounts are not satisfactory in the context of SOSR models. Building on criticisms of normative uses of SOSR models, the chapter concludes with a sketch of our account of model-based instrumental norms, which is further fleshed out in the coming chapters.

3.1 Central Characteristics of Instrumental Normativity

3.1.1 Conditional Normativity

Instrumental normativity is part of what is called *instrumental reasoning*: a kind of reasoning where, as Korsgaard [51] puts it, "the fact that a certain action is a means to your end provides you with a reason to do it" [50, pp. 37–38]. Instrumental norms are specific kinds of norms that require, permit or otherwise positively evaluate taking the means to one's ends [13, p. 5458]. An example of a instrumental norm is the following: if you have a reason to be at some place, then you have reason to travel to that place. Travelling to the place is the means used to achieve the end in question: namely, to be at that place. Notice that the normativity implicated in instrumental norms is of a conditional kind: any reason for taking the means is *conditional* on having the end in question. In this sense, instrumental normativity contrasts with what may be called *unconditional* kinds of normativity, in which some action is prescribed without further qualifications such as in the case of ethical norms that unconditionally prohibit certain classes of actions. Whilst this kind of normativity has an unrestricted scope, the normativity implicated in instrumental norms effectively lies in the prescription of taking means that facilitate one's ends.[6] In this sense, instrumental normativity does not provide a reason for an action independent of its end: the normativity is conditional on the ends given. Having a reason to be in Rotterdam tomorrow, for example, may provide reasons for hiking to Rotterdam. Although this end may provide reasons for taking the means of hiking, the reasons in question are conditional on the end: it may in fact be that there are no unconditional reasons for travelling to Rotterdam.

3.1.2 Counterfactual Dependencies

Central to instrumental norms is the presence of a counterfactual dependency between means and ends. Schematically put, a norm prescribing means m to achieve ends e assumes a counterfactual dependency along the following lines: if m were to be taken, then the realisation of e would be furthered. Leaving further details about the nature of these counterfactual dependencies to §6, we note that instrumental norms cannot prescribe means on the basis of ends without the counterfactual dependency between means and ends. Returning to our earlier example, having to be in Rotterdam provides a reason for travelling to Rotterdam only if the act of travelling would help bring about the end in question. In contrast, having to be in Rotterdam does not provide any reasons for actions that would not help further this end: those actions would not count as proper means. Although instrumental norms depend on counterfactual relations between means and ends, the content of these norms differs from mere counterfactual relations in that counterfactual relata are treated as means and ends. As a simple example, a counterfactual relationship between administering poison to someone and killing has a different content from the instrumental norm that administering poison to someone is a means for realising the end of killing them [43, pp. 9–10]. Šešelja [85] points out that the relevance of model-based counterfactual dependencies for real-world phenomena depends on the kind of possibilities that a model represents. Models range from empirical to abstract models, with differences in

[6] *Instrumentalism* is the view that all norms or normativity of some given domain can essentially be understood as involving instrumental relations between means and ends [13, pp. 5455–5456]. A full-blown instrumentalist view is susceptible to the *universality objection*, which appeals to norms that are universal in their authority and do not depend on possessing the relevant ends [13, p. 5462]. For our purposes, we do not need to adopt an instrumentalist stance: it is sufficient that instrumental relationships play a relevant role in the normative conclusions that social epistemologists aim to draw from SOSR models but it need not play an exhaustive role.

the kinds of possibilities represented. On the one hand, counterfactual dependencies found in models close to reality are likely to be relevant. Highly idealised models, on the other hand, may be too counterfactually distant to be of relevance to real-world systems [85, p. 541].

3.1.3 Instrumental Transmission

Instrumental norms state that given particular ends, there are reasons for taking particular means that help bringing about those ends. In providing end-facilitating means, instrumental norms take ends for granted: reasons in favor or against ends are brushed aside. Although reasons for ends are not part of instrumental norms, taking reasons for ends into consideration is essential because of two considerations. First of all, instrumental norms are not interesting in and of themselves. As McLaughlin puts it, "any part of a complex system that contributes to some performance of that system can be viewed as a means to performing that end" [62, p. 96]. In guiding action, some ends are prioritised over others in terms of the reasons counting in favor of those ends. Secondly, in addition to ends providing reasons for means, there is a dependence between reasons for ends and reasons for means in what is called *instrumental transmission*: a transmission of reasons for ends to reasons for means facilitating those ends [48, 47]. Raz [74]'s facilitative principle describes the idea of instrumental transmission in the following way:

> When we have an undefeated reason to take an action, we have reason to perform any one (but only one) of the possible (for us) alternative plans that facilitate its performance [74, pp. 4–5].

In other words, the facilitative principle claims that undefeated reasons for ends produce reasons for means. Having reasons for ends and instrumental transmission are not independent considerations. On the other hand, transmission is central if we want to transform reasons for particular ends into reasons for particular means facilitating those ends: without it, the instrumental norm cannot provide reasons for taking the means. On the other hand, reasons for ends determine if transmission is relevant at all: if we lack reasons for the particular end that a instrumental norm prescribes, then we may as well discard that norm. The implication is that instrumental norms are pertinent insofar as reasons for ends transmit to reasons for end-facilitating means and insofar as we, in fact, entertain reasons for those ends.

Applied to our topic, instrumental norms are relevant to the normative aim of social epistemologists only insofar as these norms are about ends that we have epistemic reason to pursue.

3.1.4 Normativity Failures

Instrumental normativity can fail in two independent ways. First of all, reasons for ends may fail to transmit to reasons for end-facilitating means in what may be called transmission failure.[7] Having a reason to be in Rotterdam tomorrow, for example, may not provide a reason for the means of hiking from Utrecht to Amsterdam and from Amsterdam to Rotterdam even though the means realise the end in question. Effectively, normativity fails because the reasons for given ends do not provide reasons for specific means, effectively a failure of Raz's facilitative principle. A second way in which instrumental normativity can fail is due to the presence of *defeaters*: independent reasons that defeat the prime facie reasons transmitted onto the means. Whilst transmission failure happen when the connection between reasons for ends and reasons for means is flawed, defeaters attack the reasons for means themselves [49]. The reason for hiking to Rotterdam may be defeated, for instance, by a further consideration to the extent that taking a train is preferable to hiking.

3.2 Normative Guidance

In §2.4, we discussed the view that normative models are models that provide normative guidance to agents. This raises the question what, exactly, normative guidance consists in. Beck and Jahn [8] provide the following description of normative guidance:

> [...] a model counts as a normative model [...] if it is taken to provide normative guidance. More specifically, a normative model provides normative guidance if it issues normative verdicts, that is, recommendation for how to act, and an agent can appeal to such verdicts in order to determine her course of action. Moreover, as we understand it, a normative model successfully exerts normative guidance if the agent's appeal to the model's verdicts

[7] Note: White [101] discusses an adjacent concept of *transmission failure* that is restricted to failures of the transmission of reasons for ends to reasons for *necessary* means.

> would be justified in the following sense: by following the model's verdicts, the agent would act in accordance with a normative reason provided by the model [8, p. 126].

In Beck and Jahn's view, models provide normative guidance to the extent to which these models output recommendations that agents can appeal to in deciding courses of action. Successful guidance, in turn, requires that the agent's appeal to the given recommendation accords with some normative reason that the model provides. Beck and Jahn employ the term *normative source* to refer to the property in virtue of which normative models are capable of exercising normative guidance [8, pp. 124, 131]. To see how these concepts operate, it is fruitful to look at a paradigm example of a device providing normative guidance, namely, normative arguments. Consider the following argument:

P_1 You ought to live healthy.

P_2 Living healthy cannot be achieved without regularly exercising.

$\therefore$ You ought to regularly exercise [8, p. 133].

In arriving at a normative conclusion, normative arguments employ a mix of descriptive and normative premises. An argument must use at least one normative premise to arrive at a normative conclusion.[8] This is, effectively, the normative source of normative arguments: without normative premises, arguments could not provide normative guidance. Successful normative guidance requires the argument's premises are sufficiently justified and that one's appeal to the argument's conclusion accords with some normative reason provided in the argument. In the example, this normative reason is given in premise P_1: a reasonable doubt about the normative import of the argument's conclusion has to point to this premise.

3.3 Normativity from Models: Three Accounts

In spite of the role that normative models play in policy-making and applied ethics, Beck and Jahn's question how models are capable of exerting normative guidance has unfortunately received "little systematic discussion" in the

[8] Note: this is a special case of Hume [41]'s is-ought gap, which holds that reasoning from pure descriptive to normative facts is fallacious.

literature [8, p. 124]. This section outlines and evaluates the three accounts that Beck and Jahn discuss: the model-argument account, the fitting account and the extending account. In all these accounts, models play a direct role in providing normative guidance: models produce direct normative verdicts that can be taken as recommendations for action. As we argue in §3.4.1, this leads to a more fundamental problem for applying these accounts to SOSR models, which at most provide an indirect kind of normative guidance.

3.3.1 The Model-Argument Account

The model-argument account holds that the normative guidance of normative models is not mysterious but instead analogous to that of a familiar normative device, namely, normative arguments. Let us revisit Beck and Jahn's example:

P_1 You ought to live healthy.

P_2 Living healthy cannot be achieved without regularly exercising.

$\therefore$ You ought to regularly exercise [8, p. 133].

In the case of a normative argument, successful normative guidance demands that the argument's premises are sufficiently justified. While descriptive premises such as P_2 require veridical justification, normative premises such as P_1 demand normative justification. As noted in §3.2, successful normative guidance demands that the premises are justified and that at least one of the argument's premises is normative. This is where the model-argument account draws the analogy: analogous to the success conditions of normative arguments, a normative model's success in providing normative guidance depends on whether the model's assumptions are sufficiently justified and if at least one of the model's assumptions is normative. Hence, the model-argument claims, the success conditions of normative models are no less puzzling than those of normative arguments [8, pp. 133–134].

As Beck and Jahn point out, however, the analogy between models and arguments breaks down in their different modes of justification. On the one hand, descriptive premises in normative arguments require veridical justification: if we are to accept a normative argument's premise, then it must be shown to be true. Descriptive assumptions in models, on the other hand, may be false: they "do not necessarily aim at correctly capturing a feature of the model's target". Given the different justification procedures of models and arguments, the analogy underlying the model-argument account becomes questionable [8, p. 135]. Beck and Jahn consider the option that although descriptive idealisations are false, they are "*close enough to the truth* or, put differently, they are *approximately true* descriptions of the agent that is targeted by the model". If this holds, then perhaps the different modes of justification may be taken to be similar enough to maintain the model-argument analogy. This, however, produces a more fundamental problem. As Beck and Jahn observe, deciding if a model's idealisations are close enough depends partially on "the aim the model sets out to serve". In the context of normative models, this then gives rise to the following problem: evaluating if a model's assumptions are close enough to the truth demands having justified beliefs about which entailments are correct normative verdicts. But if this is the case, then the model's normative guidance function is undermined: if the agent already has justified beliefs about which normative verdicts are correct, then the model's function of providing recommendations for courses of action becomes redundant [8, pp. 135–137].

3.3.2 The Fitting Account

The fitting account claims that normative models guide successfully to the extent that these models are capable of producing independently justified normative propositions. An account along these lines can be found in Roussos [81] and Beck and Jahn [8]. In Roussos's view, normative models are analogous to descriptive models: while descriptive models fit to descriptive data and are used to predict descriptive facts, normative models fit to *normative* data consisting of independently justified verdicts: prescriptions, e.g., "you should not believe contradictions", evaluations, e.g., "believing contradictions is irrational", or general facts involving normative concepts, e.g., "believing contradictions violates classical logic" [94, 81]. In turn, a normative model's predictions are used to generate normative verdicts, "something like the generation of normative claims, which are tested against the 'data' – our considered judgements". Successful guidance depends on the capability of these models to produce the correct normative verdicts [81, pp. 8, 10].[9]

[9] Given Roussos's framing of normative models as fitting data, their view faces issues independent of the general problem outlined below. In particular, a model may fit well to the normative data without picking up on normative patterns, for example, if it picks

Beck and Jahn's formulation is more general and builds on the idea that "the normative verdicts of the model are justified entirely by extrinsic considerations", in contrast with the model-argument's assumption that normative verdicts are justified by intrinsic features of the model. In their formulation, the role of normative models is to produce normative propositions that are independently justified. Similar to Roussos's view, a model exerts normative guidance to the extent that the model can "[direct] us at normative verdicts that are independently justified" [8, pp. 140–141].

This brings us to the problem with the fitting account. An answer to the normative source question should explain how it is that normative models can provide normative guidance. The fitting account, however, threatens to trivialise the guidance function of normative models. Under the fitting account, normative models are understood as models that are capable of producing correct normative verdicts. Since these verdicts are not provided with reasons to accepts those verdicts, however, the model itself does not do any normative work: we still need to provide independent justifications for the model's verdicts. But if that is so, then normative models are redundant in the guidance function: we could just obtain the normative guidance "directly from the sources that justify the normative verdicts in the first place" [8, pp. 141–142].

3.3.3 The Extending Account

In response to the issues facing the fitting account, Beck and Jahn [8] put forward the extending account: normative models guide successfully insofar as these models are able to "extend normative justification from cases in which we have firm normative verdicts to cases in which we lack those". Essentially, this view contends that normative models manage to summarise already justified normative verdicts. Subsequently, found patterns are extrapolated to novel situations in order to produce normative verdicts. Successful normative guidance is the case if a model extracts a pattern that is successfully projected onto correct normative verdicts in novel situations. According to Beck and Jahn, the extending account gives us normative reason to act in accordance with successfully guiding normative models as such models fits the pattern of our prior, already justified normative verdicts [8, pp. 124–125,

up on non-normative patterns that correlate well with normative patterns. Even though such a model may generate the correct normative verdicts, it is difficult to see where the model's normative source lies.

143, 144–145].

Prima facie, the extending account seems similar to the fitting account: in both cases, normative models are evaluated in terms of their ability of producing correct normative verdicts. In contrast to the fitting account, however, the extending account avoids trivialising the guiding function of normative models. Since normative models extract patterns from independently justified normative verdicts, we have reason to follow the model's verdicts. In contrast to the fitting account, then, the extending account succeeds to show that the guiding function of normative models is not necessarily dispensable: given their ability of extracting patterns, we can use normative models to guide us in cases of normative uncertainty [8, pp. 144–145]. This is, as Beck and Jahn admit, a relatively weak kind of guidance: the model's verdicts are, still, given without reasons supporting those verdicts. Even though the model may be successful in extrapolation, accepting the model's verdicts will still demand extrinsic normative justification [8, p. 144].

In our view, it is questionable if the extending account secures the guidance function of normative models. Beck and Jahn's concept of normative guidance requires that an agent's appeal to a model's verdicts accords with some normative reason given in the model. On the one hand, we may have reason to act in accordance with a model's verdicts in virtue of the fact that the model fits the pattern of prior, already justified normative verdicts. On the other hand, this does not entail that the model provides normative reasons for its verdicts: it could be that the model captures a projectable pattern in the normative data that is itself non-normative. To put it more strongly, the extending account can secure the guidance function if the following assumption holds: a model's success in extrapolating normative verdicts secures the model giving normative reasons for its recommendations. Since this seems a rather strong assumption, it may be more accurate to say that by their own lights, Beck and Jahn's extending account does not secure the guiding function of normative models.

3.4 Model-Based Instrumental Norms

3.4.1 Normative Guidance from SOSR Models

In §3.3.1-3.3.3, we discussed three accounts that aim to explain the normativity of models. The underlying notion of normative guidance as formulated by Beck and Jahn was a direct one: models directly output recommendation to which agents can appeal. Recalling the point from §2.4 that identical models can be interpreted as either descriptive or normative, it is clear that the models in these accounts are interpreted as normative as the models themselves output normative verdicts. But what kind of normativity should one expect from SOSR models? In a footnote, Beck and Jahn contrast their notion of normative guidance with the kind of guidance involved in instrumental reasoning:

> Note that descriptive models can, of course, also be used for normative purposes without qualifying as a normative model in our sense. For instance, given that an agent already knows where she should go, she could rely on a map she believes to be accurate in order to find her way to her destination. In such a case, however, the model seems to only facilitate her correct means-end reasoning that is sometimes relevant for normative purposes, but the model does not itself issue a normative verdict about what she should do [8, p. 126].

Beck and Jahn's description of instrumental reasoning seems to capture the kind of normativity that one should expect from SOSR models. In §2.1, we characterised SOSR models as modelling complex systems of interacting agents that allow for studying complex macro-level phenomena that emerge from micro-level behaviors. Connecting this point to the point from §2.2 that systems-oriented epistemologists judge socio-epistemic systems in terms of the system's epistemic outcomes, the plausible kind of normativity is one that prescribes socio-epistemic systems as means for furthering particular epistemic outcomes as ends. In this picture, the underlying counterfactual dependencies between means and ends rest on descriptive findings from models. Following Beck and Jahn's remarks, these descriptive findings are not normative on their own but can instead be used as input to instrumental norms. Recall, for instance, Kummerfeld and Zollman's conclusion that "[i]n some situations, actively encouraging unpopular, risky science would improve scientific outcomes" [53]. The model on its own did not give rise to this particular conclusion but suggested, at most, that risky science strategies improve scientific outcomes. Still, the descriptive finding can serve as an input, which in turn forms the basis of the instrumental norm that Kummerfeld and Zollman suggest.

Hence, the correct picture is that the guidance involved in SOSR models is of a wholly different kind than that described in Beck and Jahn's work.

Instead of producing direct normative verdicts, the role of SOSR models is to produce descriptively adequate relationships that can be used as inputs for instrumental norms. Since the contribution of SOSR models therefore lies in their descriptive import, Beck and Jahn's notion of normative guidance is implausible in our context: agents cannot appeal to normative reasons from SOSR models, which essentially play the role of a descriptive model. This picture suggests that the term *normative model* may be misleading in the context of SOSR models: although the model's descriptive verdicts are set to normative purposes, these models fail to count as normative seeing that the normative verdicts do not directly come from the models themselves.

This is a good place to make the role of instrumental norms in the normative aim explicit. In the normative aim, the epistemic outcomes of socio-epistemic systems are studied with the aim of drawing normative conclusions about the social organisation of these systems. In the first step, social epistemologists use SOSR models to find certain relationships between particular socio-epistemic systems and particular epistemic outcomes. In the next step, instrumental norms are based on descriptive findings that are discovered in the modelling endeavour. Instrumental norms play an intermediate role, in that normative conclusions about which socio-epistemic system is preferable depends on which epistemic outcomes are taken as ends. Consider, for example, the epistemic phenomenon of transient diversity as discussed in §2.2.1. Although socio-epistemic systems that foster transient diversity may enhance the epistemic outcome of a group-level convergence on the true hypothesis, normative weight for pursuing these systems depends on whether there are epistemic reasons in favor of fostering transient diversity in the first place. Hence, instrumental norms normatively inform the social organisation of epistemic communities insofar as these norms prescribe socio-epistemic systems on the basis of epistemic outcomes that we have reason to pursue.

3.4.2 Methodological Concerns for the Normative Aim

Social epistemologists who endorse the normative aim of SOSR models aim to use SOSR models for drawing normative conclusions about the organisation of real-world socio-epistemic communities. In the previous sections, we suggested that the normativity involved is of an instrumental kind, in which counterfactual dependencies serve as descriptive inputs to instrumental norms. Critics have pointed out, however, that the methodological practices of social epistemologists threatens to undermine the justification of these descriptive inputs, which as a result compromises the justification of resulting instrumental norms. Bedessem [9], for example, argues that epistemologists tend to underspecify the targets of SOSR models: "concrete policy advice [...] is tempered by the existence of inherent limits to modeling activities in social epistemology" because current models are often paired with "a fundamental lack of clarity about the exact *object* which is divided" [9, pp. 2–3]. Another criticism is given by Thicke [92], who contends that the methodological practice of epistemologists fails to secure the model-target link: "the current generation of formal models, with very exceptions, is unable to support *any* normative conclusions about science" because an evaluation of "how accurately those models represent the part of the world they seek to describe" is often missing [92, p. 322]. This link criticism is also voiced by Bedessem, who writes that epistemologists must "either pursue different aims or adopt methodologies that establish more significant links between models and their targets" and that, as of now, "the vast majority of formal models of science [...] are incapable of supporting their explicit aim to make normative claims about the organization of scientific research" [92, p. 316]. Martini and Fernández Pinto [59] emphasises this same point, point out "a missing link between models concerned with the social organization of scientific research [...] and their target systems". In particular, they argue that "[t]he usefulness or ability of [social organisation of scientific research] models to inform us (whether for normative or descriptive purposes) on the social dynamics of science is based on their ability to connect their target systems via data" [59, pp. 222, 232].

3.4.3 Towards An Account of Model-Based Instrumental Norms

The criticisms from the previous section are legitimate criticisms: basing instrumental norms on SOSR models is unjustified if a clear modelling target is lacking or if the model-target link is not sufficiently strong for the modelling purposes. Returning to our research question, this raises the question when findings from SOSR models can be used to draw instrumental norms that can inform epistemic communities. It is clear that a satisfactory account should overcome criticisms concerning the descriptive inputs to instrumental norms. Since SOSR models are agent-based models, avoiding these criticisms requires that the descriptive inputs come from models that have a clear target and adhere to verification and validation procedures that are well-established in the agent-based paradigm. As an example, the instrumental norm of encouraging risky science to improve better scientific outcomes can be based on Kummerfeld and Zollman [53]'s model just in case the model passes the right evaluation procedures with respect to the intended domain of application. Next to the procedural requirement, basing instrumental norms on descriptive inputs requires successful instrumental transmission: reasons for ends must transmit to reasons for means. Without successful transmission, the normative aim of social epistemologists cannot be achieved. Although a counterfactual relation between risky science and enhancement of scientific outcomes may be interesting, it does not on its own provide a norm that prescribes risky science as a means for enhancing scientific outcomes. Next to the procedural and transmission requirements, we add a contextual requirement that claims that argumentative context has a role to play in the justification of instrumental norms, which we specify below. In the resulting view, findings from SOSR models normatively inform the social organisation of epistemic communities with the help of instrumental norms, whose justification depends on the following three factors:

(*i*) *Model reliability*: justifying descriptive inputs of instrumental norms requires adequate verification and validation procedures.

(*ii*) *Argumentative context*: justifying descriptive inputs of instrumental norms requires the model's argumentative context.

(*iii*) *Instrumental transmission*: turning descriptive inputs into instrumental norms requires that reasons for epistemic outcomes as ends are to transmit to reasons for social systems as means facilitating those ends.

In the next chapters, we will lay out our account of instrumental norms from SOSR models by fleshing out each of these factors individually as well as addressing their own complications. In §4, we discuss the requirement of model reliability, which concerns methodological practices commonplace in the agent-based approach. In this context, we discuss both more general modelling considerations as well as challenges and complexities that pop up specifically for SOSR models. In §5, we discuss the roles of argumentative context in the justification process, identifying both essential and helpful roles of context. The essential role lies in evaluating if broader philosophical concepts such as *epistemic efficiency* are adequately translated into a model's computational constructs. In the absence of an adequate translation, descriptive inputs can fail to ground instrumental norms because of a mismatch in the philosophical concepts and computational constructs. A helpful role of argumentative context lies in the discovery of stable findings and difference-makers. Although these can be discovered with more traditional procedures, context is better equipped to discover dependencies between sets of assumptions and sets of conclusions. Third and last, §6 discusses the kind of instrumental transmission involved as well as what is required for successful instrumental transmission. In addition, we discuss ways in which reasons transmitted to means can be defeated by independent considerations, which can result in non-trivial value balances that demand philosophical reflection.

4 Model Evaluation

This chapter discusses the role of model evaluation in justifying instrumental norms. Our focus lies on three kinds of model evaluation strategies commonplace in the agent-based approach: verification, robustness analysis and empirical validation [83, 46, 31, 3]. In §4.1, we start with motivating why model evaluation is relevant to the normative aim of social epistemologists. Following this, §4.2 discusses the role of idealising assumptions in modelling and argues that the need of model evaluation results from a trade-off between idealisations and faithfulness. In turn, §4.3 discusses verification, the procedure that evaluates if a computer model is an adequate implementation of the prior conceptual model. We turn to the topic of robustness in §4.4, which is about the stability of a model's results under changes in the model's representation. Validation is considered in §4.5, which effectively aims to evaluate the link between a model and a model's target using empirical data and theories. Closing off, we argue that model evaluation does not suffice for achieving the normative aim in §4.6.

4.1 No Regulation Without Representation

In §3.4.2, we discussed some legitimate concerns about the modelling practices of social epistemologists that threaten to undermine the normative aim. Bedessem [9] pointed out that social epistemologists underspecify the intended target of their models. This is a problem for the following reason: without fixing a specific target, the application domain of any model-based normative claim becomes underspecified as well. Bedessem [9] raises this point specifically for division of labor models. In these models, the scientific resource allocation is loosely understood to be between "'theories', 'approaches', 'methods', 'research programs', 'research projects' etc.". Since these terms pick out different things in the world, it becomes difficult to ascertain the intended target of the model. As a consequence, it becomes difficult to see to what specific aspect of the division of labor model-based normative claims apply [9, p. 3]. Although underspecification of targets is a substantial methodological problem, it is not relevant to our discussion: we are specifically concerned with drawing norms from models. Repeating what we said in §2.3, constructs without a fixed intended target are not proper models in the sense relevant to our research question.

On the other hand, Thicke [92], Bedessem [9] and Martini and Fernández Pinto [59] stress the importance of a strong model-target link: an insecure model-target link makes it doubtful if model-based findings are indeed transferable to the model's target system. Since model-based instrumental norms are based on such findings, the model-target link has clear relevance to justifying these norms. A case in point is the idealisation of agents as Bayesian reasoners in network models from Zollman [106] and O'Connor and Weatherall [66]. As O'Connor and Weatherall [66] writes, such an idealisation may not adequately reflect actual scientific reasoning: "[r]eal humans, for example, are not perfect Bayesians, and many aspects go into scientists' decisions about what data to trust" [66, p. 873]. Given the unrealistic assumption of Bayesian reasoners, model-based findings about their reasoning may not transfer to groups of actual, non-Bayesian scientists. In the coming sections, we further discuss how gaps in the model-target link arise as well as what strategies are available for strengthening the model-target link.

4.2 Idealisation and the Model-Target Link

4.2.1 Between Falsities and Faithfulness

In discussing the model-target link, the first question to ask is why such a link is needed in the first place. The short answer is that the model-target connection is complicated by a commonplace feature of the modelling endeavour: *idealisation* or the use of abstractions, distortions or simplifications in models [84, 81, 10, 104, 85]. An effective entry-point into the topic of idealisation is Reiss [76]'s *explanation paradox*. Although this paradox was originally raised in the context of economics, it is highly applicable to SOSR models as well:

P_1 Models are false.[10]

P_2 Models are nevertheless explanatory.

P_3 Only true accounts can explain.

Following the logic, it appears that models are false but nevertheless, given their explanatory features, true accounts. Faced with this paradox, Reiss argues, the rational course of action is to refute the reasoning or challenge one of the paradox's statements. Since the reasoning is correct, tackling the paradox demands that we refute one of its premises [76, pp. 49–50].

One of the option is to refute statement (3): contrary to what (3) says, a model can, in fact, be both false and explanatory: the assumptions of a frictionless plane, point mass, isolated systems, fully rational agents and perfect equilibrium markets are all false but are a vital part of influential scientific models that nonetheless provide explanatory accounts of the world [22]. Effectively, refuting (3) amounts to the claim that false models can still be explanatory. This naturally brings us to idealising assumptions, which deliberately provide an abstracted, distorted or simplified picture of a target that can nevertheless explain aspects of that target [81, 79, 99]. Idealisations come at a cost: idealising assumptions bring about a mismatch between properties of the target and those of the model. In other words, some descriptions true about the model system will be false of the target system, and vice versa [81, p. 7]. Using Zollman's network model for modelling a real-world socio-epistemic system, for example, will give descriptions about group belief dynamics that will not transfer to the system in question, for the simple reason that real-world scientists are not Bayesian reasoners.

4.2.2 Virtues and Vices of Idealisation

Although idealisations bring about a mismatch between models and targets, not all model-target mismatches are necessarily bad. On the one hand, mismatches are unacceptable if the model-target relation becomes weak, for example, if idealising assumptions in the model misrepresent vital aspects of the target system [85, p. 2]. On the other hand, mismatches may be necessary if the target phenomenon is too complex: in that case, idealising assumptions

[10] As Reiss [76] notes, a model being false here just means that the model misrepresents the target in some way by making false assumptions about that target [76, pp. 49–50].

can make the target "amenable to theoretical or mathematical formulation" so that one can analyse the target [92, pp. 319–320]. This points at an important trade-off between idealising assumptions in the model and faithfulness to the target: more idealising assumptions can help with analysing a model's target but can also reduce the model's faithfulness to the target. [92, p. 322]. Šešelja [85] describes the trade-off as follows:

> The fewer components a model includes, the easier it gets to study causal dependencies between the given components. Nevertheless, such simplicity comes at a price: the model will likely end up being highly idealized, making it difficult to determine its relation to the real world. More precisely, the more idealized a model is, the harder it gets to exactly determine target phenomena it adequately represents [85, p. 538].

The aim is to find a proper balance between idealisations and faithfulness that generally secures transferability of inferences made in the model to the model's target system. Balancing these two raises the question what kind of idealisations are permissible. Weisberg [99] discusses two commonplace kinds of idealisation. *Galilean idealisations* distort a target with the aim of constructing a model that is simpler, more tractable or generally easier to analyse. [99, p. 640]. A clear example, here, is the assumption in Bayesian epistemology that degrees of belief can be represented by real-valued functions: this assumption is invoked to make the mathematics easier to work with [56, p. 141]. *Minimalist idealisations*, in contrast, misrepresent a target by constructing a model capturing only the core causal mechanisms of the target, whilst stripping away impertinent features of that system. Particularly, minimalists include only the factors that make a difference to the phenomenon under consideration [99, p. 642]. A case in point is the assumption that gas molecules do not collide, often made by theorists who apply Boyle's law to low-pressure gases. Although this assumption is strictly speaking false, low-pressure gases behave as if no collisions occur. Hence, collision of gas molecules is ignored in typical models [99, p. 643].

On the one hand, the discussion above demonstrates that the presence of false, idealising assumptions alone do not provide good reason for rejecting a model: idealising assumptions are part and parcel of the modelling endeavour. On the other hand, it is also clear that invoking idealising assumptions requires a sound justification. Galilean idealisations, for example, demand some justification as to why tractability concerns trump considerations about faithfulness. Hence, there exists a trade-off between idealising assumptions and faithfulness: the model must be “simple enough for us to understand their dynamics, but faithful enough to reality that we can use them to analyze real scientific communities” [63, p. 161]. The bottomline is that idealising assumptions ought not to undermine the model’s representational adequacy, taking the model’s purpose into account. In the literature, a number of accounts have been given of what should be considered in evaluating the model-target link. A first option is to evaluate if the model is sufficiently similar to the target in relevant respects. In this option, idealising assumptions are permissible insofar as model-target similarity is maintained [26, 98]. A second option is Parker [68]’s adequacy-for-purpose view, which effectively claims that the success of a model is evaluated with respect to that model’s purpose. In this view, permissibility of idealising assumptions depends on the model’s purpose, which can but need not be to provide an accurate and complete representation of the target [12, 56, 84, 83, 33, 56, 68]. An example is prediction: if a model’s purpose is to accurately predict a set of outcomes, then the model’s success depends on how well it can reproduce those outcomes. In this respect, faithfulness of the model to the mechanisms of its target are irrelevant in the evaluation stage.[11] The adequacy-for-purpose view provides a general framework for understanding when idealising assumptions are permissible, with representational aims as special purposes. Given this framework, the strength demanded of the model-target link depends on the model’s purposes. In §4.3-§4.5, we discuss three strategies to evaluate the strength of the model-target link: verification, robustness and empirical validation. Although the noted strategies do not exhaust all ways of evaluating the model-target link, these methodologies are commonplace in the agent-based approach and should therefore be adopted for SOSR models as specific instances of ABMs [30, 83].

4.3 Verification

The process of *verification* consists in evaluating if a formal model adequately implements the prior conceptual design: has the conceptual, informal speci-

[11] Although faithfulness to the target’s mechanisms can be helpful in the prediction task, the point here is that it does not factor into the evaluation of the model: given that the model’s purpose is just prediction, this evaluation itself is limited to the model’s predictive capacities.

fication been adequately translated into the formal model? Although model verification is not part of the model-target link itself, it is a crucial step in strengthening this link: without verification, it is unclear what is being linked to the model's target system. As a result, bad verification methodologies can undermine transferability of inferences made in the model as conclusions about the target system are usually drawn on the conceptual level rather than on the computational level. Hence, verification helps to ensure that one can rely on findings from the computer model for drawing conclusions on the conceptual level [83, 30, 60, 16]. To illustrate this point, consider Weisberg and Muldoon's landscape model. In this model, the conceptual setup is as follows: various kinds of researchers explore some epistemic landscape consisting of research approaches. The computer model formalises these notions in terms of computerised agents that move about within the span of a fixed neighborhood in some n-dimensional space. This conceptual-computational translation step also has to make choices at places where the conceptual setup underspecifies. Implementing the follower strategy in a computer model, for instance, requires some recourse option if none of the neighboring cells are visited. Notice that Weisberg and Muldoon's conclusions mentioned in §2.2.2 are framed in terms of the conceptual constructs of mavericks, followers, research strategies and epistemic progress, thus taking place on the conceptual rather than the computational level [2, p. 451]. Next to illustrating the process of model verification, Weisberg and Muldoon's model provides an example of how verification can go wrong. Alexander, Himmelreich, and Thompson [2] discovered a severe coding error in Weisberg and Muldoon's implementation of the follower strategy. Fixing this error invalidated two central results from Weisberg and Muldoon's work: mavericks perform significantly worse as well as the indirect influence that Weisberg and Muldoon attribute to mavericks on followers disappears. In other words, a mistake in the verification process effectively invalidated Weisberg and Muldoon [100]'s conceptual conclusions about mavericks, thus undermining their further normative claims about using risk-taking researchers to improve epistemic efficiency [2, p. 451].

4.4 Robustness

4.4.1 Three Kinds of Robustness Analysis

The *robustness* of a model refers to stability of a model's results under changes in the model's representation. Robustness of results is evaluated in what is called *robustness analysis*, which has two core aims. First of all, robustness analysis helps to ensure that observed results are not the artifact of auxiliary assumptions. In evaluating a model, we are always evaluating the model's core and auxiliary assumptions in conjunction. In varying auxiliary assumptions, one evaluates if the model's findings are driven by core assumptions or, instead, artifacts of auxiliary assumptions that cannot be confidently attributed to the real world [6, p. 373]. The second aim of robustness analysis is to better understand the exact conditions that bring about some observed phenomenon. On the one hand, evaluating robustness can show that a result has a general scope, if it holds under a broad scope of changes in model representation. This is, for instance, reflected in Rosenstock et al.'s observation that in many different modelling frameworks, transient diversity is conducive to learning. On the other hand, robustness checks can narrow down a given phenomenon to more specific conditions that give rise to the phenomenon, potentially even explaining why the phenomenon occurs. Here, a clear example is Rosenstock et al.'s re-evaluation of Zollman's network model, which demonstrated that the Zollman effect occurs in areas of parameter space that make the learning problem difficult [83, 92, 59, 46, 63, 54, 21, 16]. Weisberg [98] makes a clear distinction between three kinds of robustness analysis (RA), each focused on different representational aspects of models:

(*i*) *Parameter RA*: focuses on changes in parameter values of the model.

(*ii*) *Structural RA*: focuses on changes in structural features and assumptions of the model.

(*iii*) *Representational RA*: focuses on changes in the representational framework, modelling technique or modelling medium [40, 83, 98].

An example of parameter RA is given by Rosenstock et al.'s work, which studied the robustness of the Zollman effect under changes in parameters. Structural RA can be illustrated with the help of Hegselmann and Krause

[37]'s work on opinion formation, which includes analytical findings and computer simulations. Hegselmann and Krause's analytical results build on the assumption that the number of timesteps is infinite whilst their computer simulations work with a finite number of timesteps, allowing them to evaluate if their analytical results are robust under changes in the infinite timesteps assumption [37, 38]. An example of representational RA is given by Weisberg and Muldoon's model, which specifically takes the agent-based route to relax assumptions from earlier division of labor models in the spirit of Kitcher.

The kind of robustness analysis required depends on the kind of model involved. First of all, parameter RA seems necessary for typifications or theoretical abstractions that range over a number of target systems with varying parameter configurations. In case-based models, the need for parameter RA seems less stringent insofar as these models pick out a concrete target with known parameters. Rosenstock et al.'s work is a case in point of a robustness analysis of a theoretical abstraction. Second in line, structural RA is necessary for models that include controversial features or assumptions: independent of the target involved, it is vital to evaluate if outcomes are not just an artifact of those features or assumptions. This is exemplified in Weisberg and Muldoon's landscape model, which deliberately dropped Kitcher's assumption to test robustness of outcomes. Last but not least, representational RA comes into play whenever one seeks to evaluate the effect of the representational framework, technique or medium on the model findings. In this regard, Weisberg and Muldoon's switch from Kitcher's analytical approach to the agent-based framework is a good example. In all cases, however, the importance of robustness for the normative aim is clear: without robustness analysis, the justification of model-based instrumental norms is threatened due to the potential for model artifacts or contingent conditions that drive the model's findings.

4.4.2 Inductive Evidence from Robustness?

Although the main aim of robustness analysis is to evaluate stability of results across different representation choices, some researchers argue that robustness can have a further *inductive value*. In this view, the fact that a phenomenon is robust across a wide variety of models provides evidence that the phenomenon is present in the model's real-world target. Kuhlmann [52] and Weisberg [99] argue that a collection of models that gives rise to some common phenomenon and is sufficiently heterogeneous in parameters, structure and framework can provide this sort of inductive evidence [52, 99]. In defence of this view, heterogeneity can indeed play this role, for example, if it leads to inclusion of a model that captures the intended target sufficiently well. At the same time, it has to be noted that the inductive evidence delivered by robustness remains limited. In the first place, the view demands that phenomena hold across a wide variety of models, thus excluding phenomena that are true of the target but not of every model in the set. Second in line, heterogeneity of models still remains, as Sugden [90] puts it, "in the world of models": the inductive inference in play is from a smaller number of heterogeneous models to what holds of a more general class of models [90, p. 499]. In other words, heterogeneity of a set of models alone at most provides evidence that a finding holds of some common structure underlying that set. This, however, is not enough to provide inductive evidence: this requires that the collection of models captures something substantial about the target rather than about some common structure. Essentially, in the absence of empirical confirmation, robustness leaves us with "essentially free-floating models": although robustness can make the results of models more consistent, consistency alone is not a guide to truth [21, p. 331].

4.5 Empirical Validation

4.5.1 Securing the Link

In general, *validation* aims to evaluate if the link between a model and the model's target is reasonably strong with the help of empirical data and theories [83, 46, 17, 60]. Validation is not always necessary. In the case of theoretical models, for example, the point is to demonstrate that a certain phenomenon is theoretically possible. Since theoretical models typically invoke a large number of false assumptions to explore conceptual relations under ideal conditions, applying empirical validation techniques can miss the point of these models [60, pp. 3666–3667]. On the other hand, case-based models and typifications do require systematic validation, given that both aim to model concrete empirical phenomena [85, p. 10]. Hence, validation becomes necessary for the normative aim when the model is concerned with real-world phenomena: justifiably generalising normative claims from model to world depends on an adequate link between model and world [83]. The validation of a model has a number of interconnected but different stages:

(*i*) *Input validation*: calibrates the model's parameters to empirical data.

(*ii*) *Process validation*: evaluates the plausibility of the model's mechanisms.

(*iii*) *Descriptive output validation*: evaluates how well the model's outputs replicate existing data.

(*iv*) *Predictive output validation*: evaluates how well the model's outputs predict future states [46, 30].

On the one hand, the first two kinds of validation are about evaluating if the model reflects empirical data and theories about the model's target. Taking Zollman's network model as an example, the input validation step corresponds to using information from a given empirical target to set the model's parameters. Process validation, in turn, can be performed by evaluating if the belief updating mechanisms in the model are plausible according to our best theories about belief updating.[12] On the other hand, the last two kinds of validation require that the model is capable of reproducing and predicting data from its target. Note that these kinds of validation depend on the availability of enough data. Returning to the example, descriptive output validation could assess if, for example, the model is able to replicate past belief distributions of the target community. Predictive output validation, on the other hand, would predict future belief distributions of the community. Although each of $(i) - (iv)$ serves to strengthen the model-target link in its own way, this link is arguably strongest when all criteria are satisfied: a model calibrated to the empirical data with plausible inner mechanisms and a good record in replicating historical as well as future data is generally more plausible than a model satisfying only a subset of these criteria.

4.5.2 The Trouble with Data

An important concern is that model validation cannot always be achieved. Practical limitations may be in play: collecting certain data can sometimes be expensive, unethical or infeasible [16, 60]. In the beginning of the COVID-19 epidemic, for instance, data for validating epidemic models was sparse [102,

[12] Process validation can be relevant to theoretical abstractions as well. A case in point is Thoma [93]'s landscape model, which essentially modifies Weisberg and Muldoon's model with the aim of providing more plausible strategies for exploring the landscape. This is further discussed in §5.3.

pp. 222–223]. In some cases, the problem is more fundamental: it may be *impossible* to collect the right data. Consider the fact that in Zollman's model, evaluating the accuracy of scientific consensus demands access to the ground truth: namely, which hypothesis is correct. This information is usually unavailable, which makes it difficult to see how the model can be empirically validated on real-world target systems [59, 21, 106, 83]. In this respect, Martini and Fernández Pinto [59] notes that the unavailability of data may be a warning sign: perhaps the research in question cannot be done and has "to be restricted to feasible questions", in line with common scientific practices [59, p. 233]. In addition to data unavailability, there are cases where the model constructs do not translate well to real-world data. Recall the notion of epistemic or scientific significance in Weisberg and Muldoon's landscape models: what variable corresponds to epistemic significance? Harnagel [35] and Martini and Fernández Pinto [59] both advocate the use of bibliometric data for SOSR models [35, 59]. Whilst Martini and Fernández Pinto [59] advocates the use of bibliometric data to capture the social dynamic of the scientific enterprise, Harnagel [35] goes further to advocate "using the distribution of citation counts from scientific publications as a proxy for scientific significance" [35, p. 51].[13] Although this proposal does provide a measurable quantity for evaluating models, a deeper question remains: what counts as appropriate data for a vague model construct like epistemic significance?

4.6 Regulation from Representation?

In this chapter, it was argued that model evaluation plays a necessary role for achieving the normative aim of social epistemologists. Let us briefly summarise the key points. Idealising assumptions can drive a wedge between models and targets. Although model-target mismatches are not always bad, these can impact transferability of conclusions made in the model. Reducing the risk of untransferable conclusions requires strengthening the model-target link. In the agent-based approach, a number of strategies are available, which each play a vital role in achieving the normative aim. First of all, verification procedures ensure that conclusions drawn from a model can be relied

[13] On a more technical point, the use of the term *proxies* may be misleading. In statistics, a proxy is defined as a variable that correlates well with some given target variable. Calculating this correlation requires a well-defined target variable. But as introduced in Weisberg and Muldoon [100]'s work, epistemic significance is a vague construct: it is thus unclear in what sense citation counts is a proxy for epistemic significance.

on through evaluating if a computational model adequately implements the prior conceptual model. Second in line, robustness analysis sees to it that conclusions remain reliable under changes in non-vital aspects of a model. Third and last, empirical validation makes sure that one can rely that conclusions in the model world can be transferred to the model's target through strengthening the model-target link with the help of data and theories.

Seeing that model evaluation is necessary for achieving the normative aim, a fair follow-up question is if it is also sufficient for doing so. Given what we argued in §3, the answer is that it is not: the role of SOSR models is to provide descriptive relationships that can be turned into instrumental norms. In this respect, the role of model evaluation is on the level of transferability, namely, that relationships observed in SOSR models can be said to hold of the model's real-world target. Transferability, however, is not enough: model-based norms required not just descriptive relationships, but instrumental relationships in which reasons for ends provide reasons for end-facilitating means. This claim is further fleshed out in §6. Before this, §5 discusses the claim that next to model evaluation, argumentative context also plays a role in justifying descriptive relationships from SOSR models.

5 Argumentative Context

This chapter pinpoints three different kinds of ways in which the argumentative context of a model can play a role in the justification of instrumental norms. In §5.1, we lay out Aydinonat et al.'s argumentative landscape view as a response to problems of traditional model evaluation techniques. In §5.2, we argue that the justification of model-based norms depends on the specific specifications of broader philosophical concepts. Following this, §5.3 discusses a helpful role of argumentative context: namely, it can help decide if models are candidates for veridical justification and, moreover, which aspects of models require justification. In §5.4, we outline the family-of-models perspective, which studies dependencies between sets of assumptions and sets of conclusions and is thereby in a good position to discover difference-makers. Since we heavily draw on Aydinonat et al.'s view in identifying the role of argumentative context, we discuss and rebut two criticisms that arise for the claims made. First of all, §5.5 discusses the criticism that their view relativises the epistemic value of models to argumentative context. Closing off, §5.6 outlines a criticism from justification encroachment, which contends that contextual factors affect the veridical justification of models.

5.1 The Argumentative Landscape View

Aydinonat et al.'s *argumentative landscape view* claims that understanding the value and limits of models requires an analysis of models in their respective argumentative context. Aydinonat et al. present the argumentative landscape view in terms of the following three claims:

(*i*) models function as argumentative devices,

(*ii*) changes in model assumptions function as argumentative moves, and

(*iii*) drawing strong conclusions requires evaluating a family of models.

In this view, models effectively function as argumentative devices that can be invoked to analyse, attack or defend philosophical positions [6, 72]. Changes in model assumptions effectively serve as ways to explore the landscape of arguments in which different dependencies between model assumptions and model conclusions can be discovered. As an example, there is stronger evidence for a model's finding if that finding repeats itself across models with varying sets of assumptions as this decreases the likelihood that the finding is an artifact of a particular set of assumptions [6, pp. 369, 373].

Aydinonat et al.'s view offers an alternative to what may be called the received view on the epistemology of SOSR models: the view that traditional model evaluation criteria such as empirical validation exhaustively determine the epistemic value of these models [6, 83]. In particular, Aydinonat et al. motivate their view on the basis of three problems of the received view. The first problem is that a pairwise model-target comparison is not enough to evaluate a model's representational adequacy as the required level of representational adequacy of the model typically depends on the model's purpose. As an example, abstract models may invoke more idealising assumptions at the cost of representational adequacy because their purpose is to demonstrate a general claim. Second in line, Aydinonat et al. point out that a model-target comparison is not always possible. Here, abstract models are a clear case in point: these models may not correspond to any empirically observable target or collection of targets. Third in place, the focus on pairwise model-target comparisons leaves out the model's context and relation to other models. This, Aydinonat et al. suggest, conflicts with usual scientific practice in which a model's value is evaluated in the light of other models, explanations and theories [6, pp. 369, 373–374].

5.2 Models as Argumentative Devices

The first part of Aydinonat et al.'s view is to interpret models as another kind of argumentative device in the epistemologist's toolbox: models support a broader philosophical argument that involves further, implicit argumentative steps not intrinsic to those models. A clear example is given by conceptual models. As explained in §4.3, conceptual models are informal specifications, which are translated into executable computer models. As we saw, computationally implementing conceptual models involves choice as a conceptual model is typically an underspecified blueprint for its computational implementation. As an example, Weisberg and Muldoon's landscape model starts from a conceptual model of research exploration, in which scientists seek to gather epistemic significant approaches. In the computational implementation, this conceptual construct of scientists is translated to the computational construct of agents that aim to find high-elevation patches in a grid-like landscape. As we saw, in drawing conclusions about cognitive diversity, Weisberg and Muldoon move back from the computer model back to the language of the prior conceptual model, framing the simulation findings in terms of the conceptual constructs of mavericks, followers and epistemic efficiency. Given that conceptual constructs do not automatically give rise to their computational implementation and seeing that philosophical conclusions are usually drawn on the conceptual level, it is valid to ask if findings from a computational implementation genuinely hold of the initial conceptual model. Verification, to recall, is a validation technique that aims to secure that the computer model adequately implements the informal specifications of the conceptual model. This requires some kind of argumentative link establishing that particular computational constructs are, in fact, adequate implementations of the concepts specified in the prior conceptual model [6, pp. 376–379].

In addition to this, there is a further influence of argumentative context on the justification of model-based findings: the dependence of conceptual models on particular specifications of philosophical concepts. An instructive example is the concept of epistemic efficiency as employed in landscape modelling. Aydinonat et al. observe that in the literature on landscape models, the concept of epistemic efficiency has been cashed out in different ways, for example, as the amount of time progressed before a population finds the peaks of a landscape, as the proportion of discovered non-zero patches, or as the proportion of discovered landscape patches [6, pp. 380–381]. Pöyhönen

[73] suggests that none of these conceptualisations satisfactorily cash out the concept of epistemic efficiency. As a simple example, defining epistemic efficiency as the proportion of non-zero patches does track the amount of epistemically significant research approaches discovered but is insensitive to the total amount of epistemic significance of found approaches. In a similar vein, a population may take a long time before finding the peaks of a landscape but still gather overall large amounts of epistemic significance [73, p. 4525]. Since model-based findings about epistemic efficiency depend on some specification of the concept of epistemic efficiency, further philosophical conclusions from these findings hinge on how adequate that specification is. In turn, model-based instrumental norms hinge on how well philosophical concepts are cashed out. For example, a model-based norm prescribing a population of polymorphic research strategies as a means to increase epistemic efficiency is justified just in case the model's conceptual construct of epistemic efficiency is adequate enough for the purposes at hand. The upshot of this is that a second argumentative link is required to justify that philosophical concepts are specified in an adequate way. Without this link, the danger is that a mismatch occurs between instrumental norms and their descriptive inputs: the descriptive input may not ground a particular instrumental norm because it does not adequately capture central concepts in that norm.[14]

5.3 Argumentative Goals and Moves

The second part of the argumentative landscape view asserts that changes in model assumptions and the introduction of novel models are best seen as ways of attacking, analysing or defending particular philosophical positions. In clarifying this point, Aydinonat et al. distinguish *argumentative goals* from *argumentative moves*. First of all, argumentative goals concern the philosophical conclusions that philosophers aim to establish:

(*i*) demonstrating possibility or impossibility claims,

(*ii*) introducing novel ideas or considerations,

[14] In the typical case, different specifications of philosophical concepts are available. The role of argumentative context in evaluating the merits of these different specifications is crucial. In some cases, it may be that there exists no one-size-fits-all conceptualisation, which may be a sign that the concept is not fit for an actionable instrumental norm in the first place.

(*iii*) examining, establishing or challenging the validity, generality or scope of earlier argument,

(*iv*) supporting or undermining earlier claims about difference-making,

(*v*) modifying an earlier argument to correct mistakes or make the argument more plausible,

(*vi*) provide additional arguments supporting the premises or conclusions of earlier arguments, and

(*vii*) broadening the debate by introducing a new perspective on a given argument.

Notice that in achieving argumentative goals, different kinds of means can be employed: arguments, thought experiments, models and other formalisms are all legitimate means for achieving argumentative goals.[15] Second in line, argumentative moves are ways of reaching argumentative goals, such as criticising or amending earlier arguments or constructing novel arguments altogether. Aydinonat et al. describe *modelling moves* as ways in which models help to reach argumentative goals, providing the following examples:

(*i*) modifying assumptions of conceptual model,

(*ii*) implementing assumptions of conceptual models into computer models,

(*iii*) articulating informal assumptions that lead to more general philosophical conclusions, and

(*iv*) introduce novel sets of assumptions and new models.

Analogous to the argumentative moves that philosophers make in debates, each of these modelling moves helps to criticise, evaluate and explore arguments, assumptions and conclusions [6, pp. 375–376, 384, 387].

Given this picture, the influence of argumentative moves and goals can be described as follows: the value and limits of models depends on the argumentative goal that the model supports as well as the extent to which modelling

[15] Although thought experiments can be seen as special instances of arguments, this is a contentious philosophical issue. See Brown and Fehige [11] for more in-depth discussion on this issue.

moves help achieve that goal [6, pp. 375–376]. To flesh this point out, it is fruitful to consider the rise of different landscape models in the footsteps of Weisberg and Muldoon [100]'s original landscape model. Each of these models sought to challenge the validity, generality and scope of the original model by making changes in the original assumptions [6, p. 385]. Thoma [93]'s landscape model is a case in point. First in place, Thoma pointed out that the original model's Moore neighborhood assumption is unrealistic in that it depicts researchers as shortsighted and inflexible in exploring research approaches. Hence, their work introduced a variable range of movement to counter this original assumption.[16] In the second place, Thoma pointed out that mavericks and followers give an unrealistic picture of science: mavericks completely ignore work of others whilst followers completely duplicate already existing work. In response, Thoma replaced these agents with so-called explorers and extractors, who respectively take approaches very different from and very similar to already taken approaches. Contrary to Weisberg and Muldoon's conclusion that more exploration is always better, Thoma's model suggests that the division of labor is optimal when scientists are not too inflexible and not too uninformed about other lines of research [93, pp. 455, 461–464, 471]. Putting this together, Thoma's model modified assumptions from Weisberg and Muldoon's original model to make for a more realistic model, which led to different conclusions. In this respect, the value and limits of Thoma's model depend on how well the modelling moves helped towards achieving their particular goal: do the assumptions, in fact, make for a more realistic model?

A follow-up question one might ask is how argumentative goals and moves are connected with the veridical justification of a model and its findings. On the one hand, argumentative goals are relevant for deciding whether given models are candidates for veridical justification. Modelling moves promote a model's veridical justification just in case these moves work towards an argumentative goal that includes representational adequacy concerns. Hence, those argumentative goals can function as indicators for deciding if it is fruitful to evaluate a model's representational adequacy in the first place. Thoma's model was a clear case where representational adequacy was part of the argumentative goals of the model. But this need not be so: there are cases where argumentative goals and representational adequacy concerns do

[16] On a two-dimensional grid, the Moore neighborhood of a particular position is given by the agent's position and the eight cells surrounding it.

come apart. Abstract models, for example, may simply aim to demonstrate a possibility claim, which does not always require a representationally adequate model. Taking an example from the literature on belief polarisation, Singer et al. [86] developed a model in which belief polarisation emerges from epistemically rational belief-forming procedures. In introducing their model, Singer et al. emphasise that their model aims to demonstrate how belief polarisation can possibly arise from rational ways of belief formation rather than providing a realistic model that accounts for the actual causes of belief polarisation [86, pp. 2243–2245, 2265].

On the other hand, argumentative goals can play a crucial role in pinpointing the kind of veridical justification required of a model. Given the presence of idealising assumptions in models, not every aspect of a given model is a legitimate candidate for veridical justification: some aspects of the model will be false but not relevant to the model's final evaluation. In this respect, a model's argumentative goals and the moves working towards it are essential in deciding which aspects of a model demand veridical justification: it tells us what the model aims to demonstrate and which aspects of the model work towards this aim. Thoma's claim that explorers and extractors are more realistic than Weisberg and Muldoon's mavericks and followers, for example, clearly pinpoints where veridical justification should take place: empirical validation techniques can be leveraged to compare the plausibility of their behavioral mechanisms or assess how well data from real-world target systems is reproduced. Without taking argumentative goals into account, the project of veridical justification could not get off the ground: every aspect of the model would be judged equally in the evaluation stage.

5.4 The Family-of-Models Perspective

The last part of the argumentative landscape view advocates for interpreting the value and limits of models from what is called a *family-of-models perspective*. In rough terms, a family of models is a collection of models that share certain resemblances, for example through membership in the same modelling framework such as epistemic landscape models, epistemic network models or division of labor models. The family-of-models perspective does not focus on individual models but instead takes a collection of models as the unit of analysis, with the aim of mapping out the dependencies between sets of assumptions and sets of conclusions from these models. In mapping out these dependencies, the family-of-models perspective takes a more comprehensive perspective, which can help bring about a better understanding of various relationships between modelling assumptions and particular model findings. Particularly, it becomes possible to identify which modelling assumptions persistently bring about some given outcome or what assumptions are required to bring about a given outcome

A first reason for adopting the family-of-models perspective is that it can provide stronger warrant for conclusions than individual models: if a conclusion remains true under varying sets of assumptions, then it is more credible than if the conclusion comes from an individual model. One reason for this is that in the case of an individual model, results may be an artifact of the model's particular set of assumptions rather than tracking something robust. As noted in §4.4, varying assumptions is important as it can help identify if observed results are due to the core assumptions of a model. A second reason for adopting the family-of-models perspective is that the view helps with discovering *difference-makers*: factors that lead to changes in some outcome of interest, holding all other factors constant. Difference-makers relate to counterfactual outcomes in the following sense: if F is a difference-maker for G, then counterfactually, F affects G under some given ceteris paribus clause. As an example of a difference-maker, Aydinonat et al. mention the proportion of mavericks in Weisberg and Muldoon's model, which was seen to affect epistemic progress. This is a case of parameter robustness analysis: Weisberg and Muldoon examined the effect of changing the proportion of mavericks on epistemic progress. Robustness analysis, however, has limited value for discovering difference-makers. Whilst agent-based models usually consist of a collection of modelling assumptions that individually and jointly affect model findings, robustness analysis is limited to evaluating the effect of one assumption at a time. In evaluating the effect of an assumption, robustness analysis thus has a difficult time eliminating the effect of other assumptions included in the model. In contrast, the family-of-models perspective can judge the effect of multiple interventions through mapped dependencies [6, pp. 382–383, 388–391].

5.5 Is it Arguments All the Way Down?

In the previous sections, we layed out the argumentative landscape view, which construes SOSR models as argumentative devices whose function is to explore the argumentative landscape [6, 72, 84]. Politi [72] has raised a criticism of this view. According to them, if we follow Aydinonat et al.'s line of reasoning, then we must conclude that "an SOSR model is to be considered successful if it generates further arguments". This, Politi argues, gives rise to a tension between representational adequacy and achievement of the normative aim:

> Considering SOSR models as possessing a philosophical and argumentative function would rescue them from some of the criticisms about their limitations: if their function is not to provide an adequate representation of actual scientific communities, then the fact that they are not empirically validated is not a problem anymore. By doing so, however, the idea that SOSR models can be used to justify science-policy recommendations will have to be given up, or at least weakened, despite the fact that such an idea is explicitly held by some of their developers [72, p. 4].

Recapping Politi's argument, there are two options for social epistemologists. On the one hand, one can opt for the normative aim and demand that SOSR models adequately represent socio-epistemic communities. This option is prone to criticisms about the representational adequacy of SOSR models. On the other hand, SOSR models can be understood in terms of an argumentative function. Although this option escapes criticisms about model limitations, Politi's suggestion is that there exists a tension between these two options: the function of generating further arguments cuts away at the representational adequacy required for achieving the normative aim. Prima facie, this claim appears implausible: generating further arguments can go hand in hand with justifying science-policy recommendations. A trivial case of this is when the model-based arguments in question are being made about actual socio-epistemic communities. Reading between the lines, the tension that Politi identifies occurs just in case the argumentative function of models disregards adequate representation of actual socio-epistemic communities. Under this assumption, the normative aim does become more difficult to sustain, given that generating further arguments does not necessarily respect representational adequacy. Hence, the argumentative landscape view would leave the door open for models that neglect any connection with socio-epistemic communities.

The assumption that the argumentative function of SOSR models disregards adequate representation is, however, an untenable assumption. In the first place, even when modellers want to satisfy certain argumentative goals, modellers often intend to assert claims about socio-epistemic communities and consider representational adequacy concerns to achieve this. Representational concerns can even constitute a vital part of the argumentative goals of modellers. Thoma [93], for example, replaces assumptions from Weisberg and Muldoon's work with the aim of obtaining a more realistic, plausible model of the division of labor as was discussed in §5.3 [93, pp. 462–463]. Politi's claim is that in the light of inadequate model representation, modeller's intent to make recommendations for actual socio-epistemic communities "will have to be given up, or at least weakened" [72, p. 4]. This, however, follows just in case models do not adequately represent their target communities. Given that modellers do take into account representational adequacy concerns, it follows that the epistemic value of models depends on more than just the model's argumentative context.

In addition, it is unfair to suggest that Aydinonat et al. hold the view that Politi's criticism suggests. First of all, Aydinonat et al.'s motivations for the argumentative landscape view as layed out in §5.1 at most support the view that argumentative context is necessary for appreciating the value of models, not that it is sufficient for doing so. In the second place, Aydinonat et al. write that "[a]nalyses of models should include the argumentative context because their full epistemic value can only be perceived when their use is seen as part of argumentative exchange" [6, p. 392]. In their warning that ignoring argumentative context risks losing sight of the "full epistemic value" of models, Aydinonat et al. implicitly suggest that part of a model's epistemic value can be understood without considering the model's argumentative context. In other words, part of a model's epistemic value is constituted in terms of factors outside of the model's argumentative context. Last but not least, the criticism seems misguided due to the fact that Aydinonat et al. suggest that their view may apply to the sciences as well:

> Assessment of the epistemic contribution of highly simplified theoretical models is also difficult in the sciences. Modellers in the sciences may have different argumentative goals from those of their philosophical cousins, but general ideas about the importance of the argumentative context, the implicit steps between the model and the intended theoretical conclusions, the significance of argumentative and modelling moves, and the contribution of the collective exploration of model variants apply in both contexts [6, pp. 392–393].

If Aydinonat et al.'s view did claim that argumentative context is exhaustive, then this claim would have to apply to models in the sciences as well: the value of scientific models would then also be exhausted by their argumentative context. But this is a rather extreme view, which does not find a basis in Aydinonat et al.'s article.

5.6 A Threat of Pragmatic Encroachment?

In this chapter, we outlined three ways in which argumentative context are relevant to the justification of an instrumental norm. First of all, context plays an essential role in linking computational to conceptual constructs as well as evaluating particular conceptualisations of philosophical concepts. Second in place, context can play the role of an indicator in deciding if a model is a candidate for veridical justification and which aspects of the model demand veridical justification. Third and last, argumentative context can provide effective means for mapping out dependencies between sets of assumptions and sets of conclusions, which helps with identifying stable results and difference-makers. In this context, it is proper to address a potential misunderstanding that this view amounts to what is called *justification encroachment*: the view that the epistemic standards required for a belief to count as knowledge varies with practical factors [44, p. 2]. The line of reasoning would be that the relevance of context to the justification of instrumental norms amounts to the encroachment claim that the justification of these norms varies with argumentative context. This is, however, not the view defended here: the claim is that certain aspects of context missing from standard model evaluation are relevant to the *veridical* justification of instrumental norms. In this view, context plays a role in justification insofar as it can point towards aspects relevant to veridical justification. In this regard, it is not necessary to endorse encroachment theories for agreeing with our view.

6 Instrumental Transmission

In this chapter, we describe the kind of instrumental transmission involved in instrumental norms from SOSR models and outline under what circumstances transmission succeeds in the production of reasons for pursuing particular socio-epistemic systems. In §6.3, we discuss principles of instrumental transmission together with desiderata that such transmission principles should satisfy. Subsequently, §6.4 advances the claim that the desiderata discussed in §6.4 are enough to characterise how reasons for epistemic outcomes give reasons for socio-epistemic systems that help bring about those outcomes as well the strength of the reasons transmitted. In §6.5, it is argued that reasons for epistemic outcomes always produce reasons for socio-epistemic systems that help bring about those outcomes as long as the relationship underlying the instrumental norm is not descriptively false. As further illustrated in §6.6, transmitted reasons are susceptible to defeaters, which can demand a non-trivial balancing of values.

6.1 The Function of Transmission in the Normative Aim

In §3.1.4, we introduced the concepts of instrumental transmission: when reasons for ends produce reasons for means that help bring about those ends. How does instrumental transmission figure into the normative aim of social epistemologists? As an instance of systems-oriented epistemology, the agent-based approach underlying SOSR models studies socio-epistemic systems in terms of their epistemic outcomes. On the one hand, epistemologists evaluate if a system's functioning is, in fact, conducive to the system's specified epistemic ends. On the other hand, epistemologists aim to, as Goldman puts it, "identify alternative organizational structures that might be epistemically superior to the existing systems" [27, pp. 18–19]. Since systems-oriented epistemologists judge socio-epistemic systems in terms of their epistemic outcomes, identifying alternative socio-epistemic systems demands a comparison of their respective epistemic outcomes. To do so, epistemologists must be able to determine how valuable particular outcomes are as well as what epistemic reasons we have for pursuing particular outcomes. Given this picture, the role of instrumental transmission is clear: without transmission, social epistemologists cannot go from reasons for epistemic outcomes to reasons for socio-epistemic systems that facilitate those outcomes.

6.2 Normativity Failures Reconsidered

Earlier in §3.1.4, we discussed kinds of normativity failures that can occur for instrumental norms: transmission failure and defeaters. In what follows, it is insightful to tailor this discussion to the context of SOSR models specifically. Recall that in transmission failure, reasons for ends fail to transmit to reasons for means. This effectively amounts to situations in which reasons for particular epistemic outcomes fail to produce reasons for socio-epistemic systems that help bring about those outcomes. As an example, take Zollman's suggestion that in some cases, transient diversity can be improved by decreasing the flow of information. Transmission failure would be the case if reasons for transient diversity fail to provide reasons for decreasing the flow of information. In the second place, defeaters are independent reasons that defeat reasons transmitted from epistemic outcomes to socio-epistemic systems. Hence, prime facie reasons for particular socio-epistemic systems are attacked, typically by considerations about downstream effects of these systems. Here, an example is given by Rosenstock, Bruner, and O'Connor's claim that there are better solutions to Zollman's problem: the existence of better solutions undercuts reasons for socio-epistemic systems that reduce information flows.

6.3 What Kind of Transmission?

Taking on its own, the concept of transmission is ambiguous: reasons can be transmitted in different ways. In the task of disambiguating the concept of transmission, philosophers have proposed a variety of so-called *transmission principles*, each stipulating how reasons or oughts for ends e transmit to reasons or oughts for end-facilitating means m:

($\mathbf{TP}_1$) *Ought necessity*: if one ought to e, and m-ing is a necessary means to e-ing, then, because of that, one ought to m.

($\mathbf{TP}_2$) *Strong necessity*: if there is reason for one to e, and m-ing is a necessary means to e-ing, then that is a reason, at least as weighty, for one to m.

($\mathbf{TP}_3$) *Weak necessity*: if there is reason for one to e, and m-ing is a necessary means to e-ing, then that is a reason, of some weight, for one to m.

Parallel to ($\mathbf{TP}_1$)-($\mathbf{TP}_3$), philosophers have formulated sufficiency variants, in which means m are taken to be sufficient rather than necessary for bringing about ends e [47, 48].

A minimal condition on a transmission principle is that it does not trivialise instrumental normativity: reasons for ends should not produce reasons for any action whatsoever. Kolodny [47] identifies two desiderata for transmission principles. In the first place, Kolodny calls attention to a probabilisation desideratum: reasons transmit to means just in case taking the means make it more likely that the ends are brought about. Without this requirement, it would not be proper to use the term *means*: an action that does not raise the likelihood that the end comes about does not help to bring about that end. Means probabilisation comes in degrees: some means make the realisations of ends more likely than other means. Hence, the strength of reasons transmitted on means depends not just on the reasons for ends but, in addition, on how likely it is that the end comes about given that one takes the means. Kolodny formalises the probabilisation desideratum as follows:

> *Means probabilise*: if there is a reason to e and there is positive probability, conditional on one's m-ing, that one's m-ing, or some part of one's m-ing, helps to bring it about that one e's, then that is a reason to m, whose strength depends on the reason to e and on the probability [47, p. 3].

In the second place, the means must be *non-superfluous* with respect to the end: in other words, the means must play a non-redundant role in the end's realisation [47, p. 13]. Kolodny formalises the concept of superfluous means as follows:

> *Superfluous means*: means m are superfluous with respect to ends e at outcome w just in case there exists some means m^* such that
>
> (i) one m^*'s at w
>
> (ii) one's m^*-ing at w is no part of one's m-ing at w
>
> (iii) at w, if one did not m, one would still m^* and one's m^*-ing would still facilitate e

Without this desideratum, a transmission principle could produce reasons for taking means that would effectively play a redundant role in the achievement of the relevant ends [47, p. 14].

Kolodny argues that the strong transmission principles encounter significant problems connected to these two desiderata. On the one hand, strong necessity principles are problematic in that "necessary means may do little to 'probabilise' the end" [47, p. 5]. Suppose, for example, that it is necessary to take the train to be in Amsterdam. Suppose, further, that Alex and Charlie have the same reasons to be in Rotterdam and that Alex has a probability of 0.9 that the train actually arrives in Rotterdam whereas Charlie has a probability of 0.1. Although Charlie's reasons for taking the train seem less strong due to the low probability of success, strong necessity principles yield the verdict that both Alex and Charlie nonetheless have equally weighty reasons for taking the train. On the other hand, strong sufficiency principles are problematic in that "sufficient means may be 'superfluous' with respect to the end". More particularly, it can happen that in every possible outcome, the end would come about, irrespective of whether one takes the means. Kolodny illustrates this claim with an example of a pain-relieving treatment. Imagine a doctor who at every possible outcome administers a patient with Drug 1. Administering this drug will relieve the patient's pain for sure. Imagine, further, that giving the patient Drug 2 first neutralises the effects of Drug 1 but subsequently compounds with Drug 1 in the patient's bloodstream to become Drug 3, which will again surely relieve the patient's pain. Although treating the patient with Drug 2 is superfluous for relieving their pain and thus no reasons seems to transmit to taking that means, administering Drug 2 is a sufficient means for relieving the patient's pain. Hence, sufficiency principles claim that reasons nevertheless do transmit to administering Drug 2, in spite of the fact that this drug plays a redundant role in relieving the patient's pain [47, pp. 2, 13, 15].

6.4 Transmission for Model-Based Instrumental Norms

The upshot of §6.3 is that strong transmission principles are not generally tenable. Still, this raises the following question: are strong transmission principles perhaps nonetheless tenable for instrumental norms from SOSR models? Unfortunately, this seems false: Kolodny's problems equally well apply to the case of SOSR models. On the one hand, strong necessity principles turn out problematic for particular socio-epistemic systems that are necessary for some epistemic outcome but have different success probabilities depend on chance factors. Suppose, for the sake of the example, that there are fixed institutional restrictions that are necessary for avoiding the free-rider problem that Kummerfeld and Zollman [53] identify. Clearly, the success of institutional restrictions depends on various chance factors, for instance, the willingness of scientists to co-operate. Strong necessity principles will, however, yield the verdict that the reason is equally strong independent of the probability of success that those instructional restrictions have for avoiding the free-rider problem. As for sufficiency principles, it is enough to imagine some socio-epistemic system that is superfluous to a given epistemic outcome in that the outcome would have been brought about regardless. Changing the internal organisation of groups to counter belief polarisation, for example, may be unnecessary if the polarisation is due to factors that were to expire regardless.

Transmission of oughts seems too strong for instrumental norms, whose normativity is effectively of a conditional nature. Even if some socio-epistemic system facilitates an epistemic outcome that we ought to pursue, it need not be the case that we unconditionally ought to pursue that system. What is more, weak transmission principles are not very informative. Although the weak principles stipulate that reasons are transmitted, it is left open how much reason is transmitted or on what the strength of the transmitted reason depends. Hence, it seems that none of the discussed transmission principles are satisfactory for our purposes. This is, however, not a serious problem: instead of adopting a full-fledged transmission principle, taking Kolodny's two desiderata suffices for understanding the role of reasons for epistemic outcomes as well as the effectiveness of socio-epistemic systems in instrumental transmission. Taken together, instrumental transmission is understood in terms of means that probabilise ends in a non-superfluous way:

> *Means probabilise superfluously*: if there is reason for one to e, and there is positive probability, conditional on one's m-ing, that one's m-ing, or some part of one's m-ing, helps to bring it about that one e's *in a non-superfluous way*, then there is reason for one to m, whose strength depends on the reason for one to e and on the probability [47, p. 13].[17]

Returning to the discussion of counterfactual dependencies in §3.1, we noted that the counterfactual dependencies underlying instrumental norms require that taking m would further the realisation of e. We can make sense of this with the help of Kolodny's principle: that m furthers the realisation of e means that taking m makes the realisation of e more likely in a non-superfluous way. Although stronger constraints can be put on these counterfactual dependencies, Kolodny's principle gives enough material to provide

[17] Kolodny provides a final account called *general transmission*, which effectively adds a clause to deal with problems that arise from repeated applications of instrumental norms [47, p. 16]. Although this issue goes beyond the scope of the discussion, it may be that this clause needs to be added in our context as well.

an account of successful instrumental transmission that takes into account both the reasons in favor of epistemic outcomes as well as the effectiveness of socio-epistemic systems. On the one hand, the strength of reasons for particular epistemic outcomes can affect the strength of reasons in favor of particular socio-epistemic systems. Reasons for limiting communication in order to foster transient diversity, for example, depend in their strength on how strong reasons in favor of transient diversity are. On the other hand, the probability that those socio-epistemic systems help bring about the given outcomes can affect the strength of transmitted reasons. In this regard, reasons for socio-epistemic systems that limit communication becomes less weighty as the probability of those systems achieving transient diversity shrinks.

6.5 There is Such a Thing as Free Lunch

In this section, we defend the view that reasons for epistemic outcomes automatically transmit to reasons for socio-epistemic systems that help bring about those outcomes. This claim is captured in the principle below:

> *Default transmission*: reasons for epistemic outcomes automatically produce reasons for socio-epistemic systems that help bring about those outcomes.

Recall from §2.2 that systems-oriented social epistemologists work in a consequentialist framework: socio-epistemic systems are evaluated in terms of how much these systems promote or impede particular epistemic outcomes [27, p. 14]. Some examples of epistemic outcomes studied in SOSR models are speed of belief convergence and group belief accuracy [106]; epistemic significance [100]; the presence of free-riding [53]; and the extent of polarisation [66]. Systems-oriented epistemologists take outcomes and evaluate socio-epistemic systems on the basis of their effectiveness in promoting or impeding them. Given reasons for those outcomes, then, there are prima facie reasons for socio-epistemic systems that help bring about those outcomes: namely, the very fact that the socio-epistemic system play a role in promoting or impeding relevant epistemic outcomes that we already have reason to pursue. Call such reasons *effectiveness reasons*: those reasons that cite the socio-epistemic system's effectiveness in facilitating particular epistemic outcomes in a non-superfluous way. In line with Kolodny's desiderata, effectiveness reasons are transmitted only if means probabilise ends in a non-superfluous way: socio-epistemic systems must make it more likely that the outcome comes about and, in addition, those outcomes would not have come about without those systems. In addition, the weight of effectiveness reasons for pursuing particular socio-epistemic systems depends on how likely it is that the system brings about the outcome, to which we return in §6.6.

Returning to transmission failure and defeaters as discussed in §3.1.4, an important question to ask is if and how effectiveness reasons suffer from transmission failure and defeaters. In the first place, effectiveness reasons suffer from transmission failure just in case the socio-epistemic system in question would not, in fact, promote or impede the relevant epistemic outcomes. The core of the problem, here, lies in the counterfactual, descriptive relationship underlying the instrumental norm: the model-based finding does not transfer to the target system in the world. In other words, transmission failure of effectiveness reasons results from the use of relationships that are descriptively false. Note that this need not be a sign of carelessness: on any notion of justification that allows for fallibility, there will be cases in which we are justified in believing a model's false results [89]. The upshot, then, is that the adequacy of the descriptive relationships determines the success of instrumental transmission: reasons for epistemic outcomes produce reasons for socio-epistemic systems just in case there exists a counterfactual dependency from the socio-epistemic systems to the epistemic outcomes in question. As pointed out in §6.4, the strength of transmitted reasons varies with both the reasons for the outcomes as well as the system's effectiveness in non-superfluously helping to bring about the outcome. Reasons transmitted are, however, not final reasons. Even if a system has a one-hundred success rate in bringing about the outcome and the reasons for the outcome are as strong as can be, it may still happen that reasons transmitted to means are defeated by further, independent considerations that speak against the system in question. This brings us to defeaters of effectiveness reasons, which attack reasons in favor of socio-epistemic systems by calling attention to downstream effects that conflict with independent considerations. In §6.6, we discuss three notable kinds of defeaters, with the aim of illustrating that balancing effectiveness reasons and defeaters can demand non-trivial value judgments.

6.6 What Happened to Your Lunch?

6.6.1 Excessive Cost Claims

The first kind of defeater considers *excessive cost* of the means: essentially, these defeaters point out that the means considered are more costly than alternative means that are equally effective in facilitating relevant ends. In these cases, the reason transmitted to the means is, as Kolodny puts it, "typically outweighed by the excessive cost" [47, p. 15]. More minutely: reasons transmitted to the means are outweighed by the fact that there are less costly but equally effective means at one's disposal [47, 72]. Notice that balancing effectiveness reasons and excessive cost defeaters can be resolved in terms of utility calculations of effectiveness and costs. A first example can be taken from Weisberg and Muldoon's model. One way to increase epistemic success is to have a population that predominantly consists of mavericks. As Weisberg and Muldoon point out, however, it is costly to be a maverick: their risk-taking attitude puts a large burden on them. Since mixed populations may be equally effective in achieving the desired outcome, reasons in favor of maverick populations become compromised [100, pp. 250–251].

6.6.2 Individualist Epistemic Norms

The second kind of defeater claims that *individualist epistemic norms* are violated in the socio-epistemic systems to which reasons are transmitted. In essence, individualist norms may defeat effectiveness reasons because of the independence thesis that we discussed in §2.2: prescriptions of epistemic rationality between groups and individuals can diverge and even conflict. Balancing effectiveness reasons and individualist norms is a non-trivial endeavour, as it involves philosophical reflection about which of the considerations should be valued more. As an example of violations with individualist epistemic norm, consider Rosenstock, Bruner, and O'Connor's criticism of Zollman's recommendation to limit information. According to Rosenstock, Bruner, and O'Connor, Zollman's recommendation conflicts with the individualist norm that in forming beliefs, individuals should take into account all available evidence: if information is limited, then it becomes infeasible for individuals to consider all available evidence [80, p. 251].[18]

[18] As we saw in §2.2.1, Rosenstock, Bruner, and O'Connor in addition argue that there are better solutions to the problem, such as setting up scientific data standards. This move may be taken to provide an excessive cost defeater: there are equally effective means

6.6.3 Moral Defeaters

The third and last defeater is what we call a *moral defeater*. This defeater attacks reasons transmitted to socio-epistemic systems on the basis of moral considerations relating to those systems. In this context, *moral* is liberally understood as having both a descriptive and normative sense. Descriptively, moral defeaters concerns codes of conduct actually followed in some society, such as a state's law or unwritten rules of social conduct. In the normative sense, moral defeaters relate to codes of conduct that are rational to follow under specified conditions [25]. In both cases, moral defeaters involve a non-trivial balancing between effectiveness reasons and moral considerations. As an example of a moral defeater, consider Kummerfeld and Zollman's conclusion that the scientific state of nature gives rise to a free-rider problem. Although this transmit reasons to socio-epistemic systems which impose institutional control to avoid free-riding, some of these systems may collide with moral concerns about institutional control. Resnik [78] discusses science's right to self-governance and different ways in which institutional control can help or block scientific progress. Resnik advocates the view that institutional control becomes an impediment to scientific progress if it controls both the content of science as well as scientific communication [78, pp. 220, 228, 238]. Next to pointing at the consequences of institutional control, one could also take the stance that forms of institutional control are intrinsically bad in that the scientific enterprise should be clear of external forces, which could defeat transmitted reasons. An insightful overview of the various positions within this debate is given by Resnik [77].

7 Conclusion

The aim of this thesis was to answer the following question: how can findings from SOSR models normatively inform the organisation of real-world socio-epistemic systems? In sharpening this question, we identified the normativity expected from SOSR models to be a kind of instrumental normativity, in which reasons for ends produce reasons for end-facilitating means. On the one hand, the account set out in this thesis claimed that model-based findings can normatively inform the organisation of socio-epistemic systems through model-based instrumental norms. On the other hand, it explained

available to achieve Zollman's ends.

the justification of model-based instrumental norms in terms of three factors: model evaluation, argumentative context, and instrumental transmission.

First in place, model evaluation is central to justifying model-based norms because these norms depend on counterfactual dependencies that are descriptively adequate. model evaluation can help with ensuring descriptive adequacy. We identified three techniques from the agent-based literature: verification, robustness and empirical validation. Verification ensures that the computer model actually implements the prior conceptual model. Since conclusions from results in the computer model are typically drawn on the conceptual level, verification is necessary to ensure that one can rely on results from the computer model in drawing conclusions. Robustness analysis helps to ensure that a model's conclusions are stable under changes in a model's representation, rather than artifacts of idealising assumptions. Empirical validation evaluates the model-target link with the help of empirical data and theories, with the aim of securing that a model's conclusions can be safely transferred to the model's target system.

Second in line, we discussed the role of argumentative context in the justification of model-based norms. In discussing this factor, we singled out an essential role as well as two helpful roles of argumentative context. The essential role of argumentative context consists in evaluating if prior conceptual models adequately cash out philosophical concepts mentioned in instrumental norms. In this regard, the concept of epistemic efficiency was a main example: a model-based norm prescribing a socio-epistemic system to further epistemic efficiency is justified only if the model's conceptual construct of epistemic efficiency is acceptable. The first helpful role of context draws on the context of argumentative debates to single out candidates for veridical justification and, moreover, to evaluate which aspects of models require veridical justification. The second helpful role of context came to the fore within the family-of-models perspective, which lays out dependencies between modelling assumptions and conclusions, which in turn helps with the identification of counterfactual dependencies through difference-makers.

Third but not least, we analysed the factor of instrumental transmission, which is the essential ingredient for turning model-based descriptive relationships into actionable instrumental norms. Drawing on the system-oriented perspective's evaluation of socio-epistemic systems in terms of epistemic outcomes, the transmission was specifically framed as occurring from reasons for epistemic outcomes to reasons for socio-epistemic systems that help bring about those outcomes. On reviewing the literature on different kinds of transmission, we adopted a minimalist stance according to which taking the means probabilise the ends in a non-superfluous fashion. In spite of this minimalist stance, informative conclusions were drawn. First of all, transmitted reasons in favor of socio-epistemic systems depend on two things: the transmitter reason for the epistemic outcome and the probability that taking the means bring about the end. In addition, we contended that there exists default transmission in the form of effectiveness reasons that favor a socio-epistemic system in its effectiveness of facilitating a certain epistemic outcome, understood in terms of means probabilisation and non-superfluidity. Although prima facie reasons are secured as long as the descriptive relationship underlying instrumental norms is correct, transmitted reasons can be defeated, which can result in a non-trivial weighing of considerations. In this regard, we discussed three examples: excessive cost claims, individualist epistemic norms and moral defeaters.

Let us zoom out to a wider perspective on the normative aim and what successful achievement of this aim involves. The normative aim takes the opportunities of SOSR models for explaining, predicting and understanding socio-epistemic phenomena and exploits their findings from these models for redesigning, restructuring, and reorganising real-world socio-epistemic systems. According to the instrumentalist stance advocated in this thesis, the normative aim is to be understood in terms of model-based instrumental norms, which produce reasons for particular socio-epistemic systems on the basis of epistemic outcomes of these systems. The resulting view identifies model evaluation, argumentative context and instrumental transmission as vital factors for justifying instrumental norms on SOSR models.

The instrumentalist view defended has three notable consequences for achievement of the normative aim. First in line, the view implies that the normative aim cannot be achieved without achievement of the epistemic aim: constructing instrumental norms requires using SOSR models to discover counterfactual dependencies, which serve as descriptive inputs to instrumental norms. Second in place, the view suggests that exclusively empirical methods fall short of achieving the normative aim. Next to empirical methods, philosophical reflection is required to evaluate how well particular concepts are cashed out in models or for identifying epistemic or moral defeaters of transmitted reasons. Third in line, the view also indicates that exclusively philosophical methods are insufficient for achieving the normative aim. In particular, veridical justification of SOSR models should use methodologies that more generally apply to models falling under the agent-based approach.

The upshot of the view defended is that, in fact, the combination of empirical and philosophical approaches is not enough either: the success of Kitcher's aim of changing social structure for the advancement of learning depends on a wide collaboration of epistemologists, modellers, moral philosophers, policy-makers and the like, each playing their role in ensuring that reasons for particular epistemic outcomes produce undefeated reasons for the socio-epistemic order that is anticipated to facilitate them.

References

[1] Mohsen Agha-Hoseinali-Shirazi et al. "Application of Agent-Based Modeling in Water Resources Management and Planning". In: *Essential Tools for Water Resources Analysis, Planning, and Management*. Springer, 2021, pp. 177–216.

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

[3] David Anzola. "Social Epistemology and Validation in Agent-Based Social Simulation". In: *Philosophy & Technology* 34.4 (2021), pp. 1333–1361.

[4] Shahar Avin. "Policy Considerations for Random Allocation of Research Funds". In: *RT. A Journal on Research Policy and Evaluation* 6.1 (2018).

[5] Kamil Aybuğa and Aysel Gamze Yücel Işildar. "Agent-Based Approach on Water Resources Management: A Modified Systematic Review". In: *Turkish journal of water scince and management* (2022). DOI: 10.31807/tjwsm.1123808.

[6] N. Emrah Aydinonat, Samuli Reijula, and Petri Ylikoski. "Argumentative Landscapes: the Function of Models in Social Epistemology". In: *Synthese* 199.1-2 (2021), pp. 369–395.

[7] Venkatesh Bala and Sanjeev Goyal. "Learning from Neighbours". In: *The Review of Economic Studies* 65.3 (1998), pp. 595–621.

[8] Lukas Beck and Marcel Jahn. "Normative Models and Their Success". In: *Philosophy of the Social Sciences* 51.2 (2021), pp. 123–150.

[9] Baptiste Bedessem. "The Division of Cognitive Labor: Two Missing Dimensions of the Debate". In: *European Journal for Philosophy of Science* 9 (2019), pp. 1–16.

[10] Claus Beisbart. "How Can Computer Simulations Produce New Knowledge?" In: *European Journal for Philosophy of Science* 2.3 (2012), pp. 395–434.

[11] James Robert Brown and Yiftach Fehige. “Thought Experiments”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2023. Metaphysics Research Lab, Stanford University, 2023.

[12] Mark Colyvan. “Idealisations in Normative Models”. In: *Synthese* 190.8 (2013), pp. 1337–1350.

[13] Charles Côté-Bouchard. “Two Types of Epistemic Instrumentalism”. In: *Synthese* 198.6 (2021), pp. 5455–5475.

[14] Herbert Dawid, Giorgio Fagiolo, et al. “Agent-Based Models for Economic Policy Design: Introduction to the Special Issue”. In: *Journal of Economic Behavior & Organization* 67.2 (2008), pp. 351–354.

[15] Igor Douven. “Computational Models in Social Epistemology”. In: *The Routledge Handbook of Social Epistemology*. Routledge, 2019, pp. 457–465.

[16] Igor Douven. “Pandemics and Flexible Lockdowns: In Praise of Agent-Based Modeling”. In: *European Journal for Philosophy of Science* 13.3 (2023), p. 35.

[17] Igor Douven and Christoph Kelp. “Truth Approximation, Social Epistemology, and Opinion Dynamics”. In: *Erkenntnis* 75 (2011), pp. 271–283.

[18] Bruce Edmonds and Scott Moss. “From KISS to KIDS–an ‘Anti-Simplistic’ Modelling Approach”. In: *International Workshop on Multi-Agent Systems and Agent-Based Simulation*. Springer. 2004, pp. 130–144.

[19] Bruce Edmonds et al. “Simulating the Social Processes of Science”. In: *Journal of Artificial Societies and Social Simulation* 14 (Oct. 2011). DOI: 10.18564/jasss.1842.

[20] Joshua M. Epstein. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press, 2012.

[21] Daniel Frey and Dunja Šešelja. “What is the Epistemic Function of Highly Idealized Agent-Based Models of Scientific Inquiry?” In: *Philosophy of the Social Sciences* 48.4 (2018), pp. 407–433.

[22] Roman Frigg and Stephan Hartmann. "Models in Science". In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Spring 2020. Metaphysics Research Lab, Stanford University, 2020.

[23] Roman Frigg, James Nguyen, et al. *Modelling Nature: An Opinionated Introduction to Scientific Representation.* Springer, 2020.

[24] Jonathan Fuller. "What are the COVID-19 Models Modeling (Philosophically Speaking)?" In: *History and Philosophy of the Life Sciences* 43 (2021), pp. 1–5.

[25] Bernard Gert and Joshua Gert. "The Definition of Morality". In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta. Fall 2020. Metaphysics Research Lab, Stanford University, 2020.

[26] Ronald N. Giere. *Explaining Science: A Cognitive Approach.* University of Chicago Press, 1988.

[27] Alvin I. Goldman. "A Guide to Social Epistemology". In: *Social Epistemology: Essential readings* (2011), pp. 11–37.

[28] Alvin I. Goldman. *Why Social Epistemology is Real Epistemology.* na, 2010.

[29] Alvin I. Goldman and Matthew McGrath. *Epistemology: A Contemporary Introduction.* Ed. by Matthew McGrath. New York: Oxford University Press, 2014.

[30] Claudius Gräbner. "How to Relate Models to Reality". In: *An Epistemological Framework for the Validation and Verification of Computational Models* 21.3 (2018).

[31] Patrick Grim. "Modeling Epistemology: Examples and Analysis in Computational Philosophy of Science". In: *2019 Spring Simulation Conference (SpringSim).* IEEE. 2019, pp. 1–12.

[32] Patrick Grim and Daniel Singer. "Computational Philosophy". In: *The Stanford Encyclopedia of Philosophy.* Ed. by Edward N. Zalta and Uri Nodelman. Fall 2022. Metaphysics Research Lab, Stanford University, 2022.

[33] Sven Ove Hansson. "Formalization in Philosophy". In: *Bulletin of Symbolic Logic* 6.2 (2000), pp. 162–175.

[34] M. Hare and Peter Deadman. “Further Towards a Taxonomy of Agent-Based Simulation Models in Environmental Management”. In: *Mathematics and computers in simulation* 64.1 (2004), pp. 25–40.

[35] Audrey Harnagel. “A Mid-Level Approach to Modeling Scientific Communities”. In: *Studies in History and Philosophy of Science Part A* 76 (2019), pp. 49–59.

[36] Daniel M. Hausman. “Philosophy of Economics”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University, 2021.

[37] Rainer Hegselmann and Ulrich Krause. “Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation”. In: *Journal of Artificial Societies and Social Simulation* 5.3 (2002), p. 33.

[38] Rainer Hegselmann, Ulrich Krause, et al. “Truth and Cognitive Division of Labor: First Steps Towards a Computer Aided Social Epistemology”. In: *Journal of Artificial Societies and Social Simulation* 9.3 (2006), p. 10.

[39] Lu Hong and Scott E. Page. “Problem Solving by Heterogeneous Agents”. In: *Journal of Economic Theory* 97.1 (2001), pp. 123–163.

[40] Wybo Houkes, Dunja Šešelja, and Krist Vaesen. *Robustness analysis*. 2023. URL: https://philsci-archive.pitt.edu/22010/.

[41] David Hume. *A Treatise of Human Nature*. Oxford University Press, 2000.

[42] Elizabeth Hunter, Brian Mac Namee, and John D. Kelleher. “A Comparison of Agent-Based Models and Equation-Based Models for Infectious Disease Epidemiology”. In: (2018).

[43] Hans Kelsen. *General Theory of Norms*. Oxford University Press, 1991.

[44] Brian Kim. “Pragmatic Encroachment in Epistemology”. In: *Philosophy Compass* 12.5 (2017), e12415.

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

[46] Dominik Klein, Johannes Marx, and Kai Fischbach. “Agent-Based Modeling in Social Science, History, and Philosophy. An Introduction”. In: *Historical Social Research/Historische Sozialforschung* 43.1 (163 (2018), pp. 7–27.

[47] Niko Kolodny. “Instrumental Reasons”. In: *The Oxford Handbook of Reasons and Normativity*. Ed. by Daniel Star. Oxford University Press, 2018.

[48] Niko Kolodny and John Brunero. “Instrumental Rationality”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2023. Metaphysics Research Lab, Stanford University, 2023.

[49] Robert Koons. “Defeasible Reasoning”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2022. Metaphysics Research Lab, Stanford University, 2022.

[50] Christine M. Korsgaard. “The Normativity of Instrumental Reason”. In: *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. Oxford University Press, Oct. 2008. ISBN: 9780199552733. DOI: 10.1093/acprof:oso/9780199552733.003.0002. eprint: https://academic.oup.com/book/0/chapter/147550701/chapter-ag-pdf/44996279/book_5036_section_147550701.ag.pdf. URL: https://doi.org/10.1093/acprof:oso/9780199552733.003.0002.

[51] Christine M. Korsgaard. *The Sources of Normativity*. Cambridge University Press, 1996.

[52] Meinard Kuhlmann. “On the Exploratory Function of Agent-Based Modeling”. In: *Perspectives on Science* 29.4 (2021), pp. 510–536.

[53] Erich Kummerfeld and Kevin J.S. Zollman. “Conservatism and the Scientific State of Nature”. In: *The British Journal for the Philosophy of Science* (2016).

[54] Jaakko Kuorikoski, Aki Lehtinen, and Caterina Marchionni. “Robustness Analysis Disclaimer: Please Read the Manual Before Use!” In: *Biology & Philosophy* 27 (2012), pp. 891–902.

[55] Blake LeBaron and Peter Winker. “Introduction to the Special Issue on Agent-Based Models for Economic Policy Advice”. In: *Jahrbücher für Nationalökonomie und Statistik* 228.2-3 (2008), pp. 141–148.

[56] Matheus de Lima Rui. “What is the Aim of Models in Formal Epistemology?” In: *Principia: an International Journal of Epistemology* 26.1 (2022), pp. 135–152.

[57] Christian List and Philip Pettit. “Aggregating Sets of Judgments: An Impossibility Result”. In: *Economics & Philosophy* 18.1 (2002), pp. 89–110.

[58] Charles M. Macal and Michael J. North. “Tutorial on Agent-Based Modeling and Simulation”. In: *Proceedings of the Winter Simulation Conference, 2005.* IEEE. 2005, 14–pp.

[59] Carlo Martini and Manuela Fernández Pinto. “Modeling the Social Organization of Science: Chasing Complexity Through Simulations”. In: *European Journal for Philosophy of Science* 7 (2017), pp. 221–238.

[60] Conor Mayo-Wilson and Kevin J.S. Zollman. “The Computational Philosophy: Simulation as a Core Philosophical Method”. In: *Synthese* (2021), pp. 1–27.

[61] Conor Mayo-Wilson, Kevin J.S. Zollman, and David Danks. “The Independence Thesis: When Individual and Social Epistemology Diverge”. In: *Philosophy of Science* 78.4 (2011), pp. 653–677.

[62] Peter McLaughlin. “Functions and Norms”. In: *Functions in Biological and Artificial Worlds: Comparative Philosophical Perspectives* (Jan. 2009), pp. 93–102. DOI: 10.7551/mitpress/9780262113212.003.0006.

[63] Ryan Muldoon and Michael Weisberg. “Robustness and Idealization in Models of Cognitive Labor”. In: *Synthese* 183 (2011), pp. 161–174.

[64] Muaz Niazi and Amir Hussain. “Agent-Based Computing from Multi-Agent Systems to Agent-Based Models: a Visual Survey”. In: *Scientometrics* 89.2 (2011), pp. 479–499.

[65] Cailin O’Connor, Sanford Goldberg, and Alvin Goldman. “Social Epistemology”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2023. Metaphysics Research Lab, Stanford University, 2023.

[66] Cailin O’Connor and James Owen Weatherall. “Scientific Polarization”. In: *European Journal for Philosophy of Science* 8 (2018), pp. 855–875.

[67] Romans Pancs and Nicolaas J. Vriend. “Schelling’s Spatial Proximity Model of Segregation Revisited”. In: *Journal of Public Economics* 91.1-2 (2007), pp. 1–24.

[68] Wendy S. Parker. “Model Evaluation: An Adequacy-for-Purpose View”. In: *Philosophy of Science* 87.3 (2020), pp. 457–477. DOI: 10.1086/708691.

[69] Nicolas Payette. “Agent-Based Models of Science”. In: *Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences*. Springer, 2011, pp. 127–157.

[70] Eugenio Petrovich and Marco Viola. “Social Epistemology at Work: from Philosophical Theory to Policy Advice”. In: *RT. A Journal on Research Policy and Evaluation* 6.1 (2018).

[71] Manuela Fernández Pinto and Daniel Fernández Pinto. “Epistemic Landscapes Reloaded: An Examination of Agent-Based Models in Social Epistemology”. In: *Historical Social Research/Historische Sozialforschung* 43.1 (163 (2018), pp. 48–71.

[72] Vincenzo Politi. “Formal Models of the Scientific Community and the Value-Ladenness of Science”. In: *European Journal for Philosophy of Science* 11.4 (2021), p. 97.

[73] Samuli Pöyhönen. “Value of Cognitive Diversity in Science”. In: *Synthese* 194.11 (2017), pp. 4519–4540.

[74] Joseph Raz. “The Myth of Instrumental Rationality”. In: *Journal of Ethics & Social Philosophy* 1 (2005), p. 1.

[75] Samuli Reijula and Jaakko Kuorikoski. “Modeling Epistemic Communities”. In: *The Routledge Handbook of Social Epistemology*. Ed. by Fricker, Miranda and Graham, Peter J. and Henderson, David and Pedersen, Nikolaj J.L.L. New York: Routledge, 2020. Chap. 24, pp. 240–249.

[76] Julian Reiss. “The Explanation Paradox”. In: *Journal of Economic Methodology* 19.1 (2012), pp. 43–62.

[77] David B. Resnik. *Playing Politics with Science: Balancing Scientific Independence and Government Oversight*. Oxford University Press, 2009.

[78] David B. Resnik. “Scientific Autonomy and Public Oversight”. In: *Episteme* 5.2 (2008), pp. 220–238.

[79] Alexander Reutlinger, Dominik Hangleiter, and Stephan Hartmann. “Understanding (With) Toy Models”. In: *The British Journal for the Philosophy of Science* (2018).

[80] Sarita Rosenstock, Justin Bruner, and Cailin O’Connor. “In Epistemic Networks, Is Less Really More?” In: *Philosophy of Science* 84.2 (2017), pp. 234–252.

[81] Joe Roussos. “Normative Formal Epistemology as Modelling”. In: *The British Journal for the Philosophy of Science* (2021). URL: https://api.semanticscholar.org/CorpusID:245048068.

[82] Thomas C. Schelling. “Dynamic Models of Segregation”. In: *Journal of Mathematical Sociology* 1.2 (1971), pp. 143–186.

[83] Dunja Šešelja. “Agent-Based Modeling in the Philosophy of Science”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2023. Metaphysics Research Lab, Stanford University, 2023.

[84] Dunja Šešelja. “Agent-Based Models of Scientific Interaction”. In: *Philosophy Compass* 17.7 (2022), e12855.

[85] Dunja Šešelja. “Exploring Scientific Inquiry via Agent-Based Modelling”. In: *Perspectives on Science* 29.4 (2021), pp. 537–557.

[86] Daniel J. Singer et al. “Rational Social and Political Polarization”. In: *Philosophical Studies* 176 (2019), pp. 2243–2267.

[87] Abhinav Singh, Dmitri Vainchtein, and Howard Weiss. “Schelling’s Segregation Model: Parameters, Scaling, and Aggregation”. In: *Demographic Research* 21 (2009), pp. 341–366.

[88] Flaminio Squazzoni, R. Boero, et al. “Does Empirical Embeddedness Matter? Methodological Issues on Agent-Based Models for Analytical Social Science”. In: *JASSS* 8 (2005).

[89] Matthias Steup and Ram Neta. “Epistemology”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Spring 2024. Metaphysics Research Lab, Stanford University, 2024.

[90] Robert Sugden. "Credible Worlds: the Status of Theoretical Models in Economics". In: *Journal of Economic Methodology* 7.1 (2000), pp. 1–31.

[91] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[92] Michael Thicke. "Evaluating Formal Models of Science". In: *Journal for General Philosophy of Science* 51.2 (2020), pp. 315–335.

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

[94] Michael G. Titelbaum. *Normative Modeling*. Feb. 2021. URL: https://philsci-archive.pitt.edu/18670/.

[95] Philippe Van Basshuysen and Lucie White. "Were Lockdowns Justified? A Return to the Facts and Evidence". In: *Kennedy Institute of Ethics Journal* 31.4 (2021), pp. 405–428.

[96] Ben Vermeulen and Andreas Pyka. "Agent-Based Modeling for Decision Making in Economics Under Uncertainty". In: *Economics* 10.1 (2016), p. 20160006.

[97] Jonathan Weisberg. "Formal Epistemology". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2021. Metaphysics Research Lab, Stanford University, 2021.

[98] Michael Weisberg. *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press, 2012.

[99] Michael Weisberg. "Three Kinds of Idealization". In: *The Journal of Philosophy* 104.12 (2007), pp. 639–659.

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

[101] Stephen J. White. "Transmission Failures". In: *Ethics* 127.3 (2017), pp. 719–732.

[102] Eric Winsberg, Jason Brennan, and Chris W. Surprenant. "How Government Leaders Violated their Epistemic Duties During the SARS-CoV-2 Crisis". In: *Kennedy Institute of Ethics Journal* 30.3 (2020), pp. 215–242.

[103] Eric Winsberg and Stephanie Harvard. "Purposes and Duties in Scientific Modelling". In: *J Epidemiol Community Health* 76.5 (2022), pp. 512–517.

[104] Petri Ylikoski and N. Emrah Aydinonat. "Understanding with Theoretical Models". In: *Journal of Economic Methodology* 21.1 (2014), pp. 19–36.

[105] Kevin J.S. Zollman. "Network Epistemology: Communication in Epistemic Communities". In: *Philosophy Compass* 8.1 (2013), pp. 15–27.

[106] Kevin J.S. Zollman. "The Communication Structure of Epistemic Communities". In: *Philosophy of Science* 74.5 (2007), pp. 574–587.

[107] Kevin J.S. Zollman. "The Epistemic Benefit of Transient Diversity". In: *Erkenntnis* 72.1 (2010), pp. 17–35.