Forthcoming in *Synthese*

External Representations and Scientific Understanding

Jaakko Kuorikoski

jaakko.kuorikoski@helsinki.fi

Department of Political and Economic Studies

P.O. Box 24

00014 University of Helsinki

Finland

&

Petri Ylikoski

petri.ylikoski@helsinki.fi

Department of Social Research

P.O. Box 18

00014 University of Helsinki

Finland

Abstract

This paper provides an inferentialist account of model-based understanding by combining a counterfactual account of explanation and an inferentialist account of representation with a view of modeling as extended cognition. This account makes it understandable how the manipulation of surrogate systems like models can provide genuinely new empirical understanding about the world. Similarly, the account provides an answer to the question how models, that always incorporate assumptions that are literally untrue of the model target, can still provide factive explanations. Finally, the paper shows how the contrastive counterfactual theory of explanation can provide tools for assessing the explanatory power of models.

1. Introduction

The importance of model-based reasoning in science has not gone unnoticed by philosophers, and the autonomy and perceived unrealisticness of most models have raised questions concerning the way in which they can provide understanding of the world. This puzzlement can be summed up in two questions. First, how can the manipulation of these surrogate systems provide genuinely new empirical understanding about the world? Second, how can models, which always incorporate assumptions that are literally untrue of the model target, provide explanations, if explanation is taken to be factive? In addition, a viable theory of model-based explanations must be able to explicate what makes the difference between truly explanatory and merely phenomenological models. Furthermore, the evaluation of the explanatory qualities of models is usually based on rather unarticulated intuitions about explanatory goodness. An adequate account of model-based explanations should also be able to say which properties of models are the basis for these judgements and provide means to discuss their epistemic legitimacy. In this paper we show how these challenges can be met with an inferential conception of understanding. The basic elements of our account have been developed in some of our earlier papers (Kuorikoski 2011, Kuorikoski & Lehtinen 2009, Ylikoski 2013, 2014, Ylikoski & Aydinonat 2014), but this paper brings these ideas together and expands them into a systematic account of model-based explanation. We argue that approaching model-based reasoning as extended cognition and adopting an inferentialist analysis of representation dissolves both of the two central questions, and shows that the sense of paradox associated with model-based explanations arises from faulty philosophical assumptions. We also show how the contrastive counterfactual theory of explanation, on which our inferential conception of understanding is based, can provide tools for assessing the explanatory power of models.

Many things are called models and models are being used to achieve many things. The word 'model' can refer to anything from a physical scale model to a set of equations on paper and a program running on a computer. Sometimes models are taken to be the abstract entities that these various material things somehow realize or instantiate. When taken to be abstract entities, some believe models to be mathematical (set theoretic) structures or trajectories in some state space, whereas others take them to be

propositionally structured, as akin to fiction. In this paper we approach models as material objects that can be used to keep score of inferential moves and in this way be used to help in reasoning about the phenomena of interest. This amounts to approaching modeling as extended cognition. This perspective downplays two common intuitions that we find problematic: First, that scientific understanding should be conceptualized in terms of what happens in the minds of individual scientists, and second, that there is an epistemic puzzle concerning how we gain empirical understanding about the real world by examining abstract objects.[1]

Scientific models also have many functions. Data models serve to represent a mass of data in an orderly and economic fashion. Closely related phenomenological models only try to capture the salient observable regularities of the target phenomenon. Models that capture the most important patterns are good for prediction, but may not have much explanatory import. Scale models are used to simulate the behavior of the target system in conditions of interest, quite possibly with no goal of understanding why it behaves as it does. Many models are only for illustrative and pedagogical uses. In this paper we focus on the understanding provided by abstract theoretical models which need not be tested against or estimated from any specific body of data. To illustrate our claims we use the family of models commonly known as the Schelling segregation model. Nevertheless, our basic stance is applicable to the explanatory import of more complex computational models and models fitted or built to account for a specific empirical phenomenon.

The structure of the paper is as follows. We first provide the outline of our conceptions of explanation and understanding as inferential ability. In the following sections we propose to approach modeling as extended cognition, distinguish between understanding the model and understanding with the model, and show how the philosophical puzzles related to model-based explanations can be solved by supplementing this viewpoint with an inferentialist account of representation. Finally, we apply our theory of the

[1] This should not be taken as a stance in the ontology of models discussion. We do not deny that it makes sense to abstract away from the concrete instantiations of these artifacts to their inferential properties and define the identity in these terms (see Kuorikoski and Lehtinen 2009).

dimensions of explanatory power to model-based explanations. We aim to present an overall perspective on model-based understanding, and for this reason we cannot provide a detailed exposition of all aspects of our account. At these points we direct the reader to other publications in which the specifics are developed in more detail.

2. Explanation and understanding

To make sense of the explanatory use of models, we will employ an inferential account of understanding based on a contrastive-counterfactual account of explanation. The underlying motivation of our account is realistic: The aim of science is to learn about phenomena in the world and particularly the dependencies between them. These dependencies are objective in the sense that they are independent of our ways of perceiving, conceptualizing and theorizing about them. Metaphysically, we are pluralists about dependence: apart from causal dependencies, there are also constitutive and possibly formal ones, and many combinations of these (Ylikoski 2013). Knowledge of these dependencies constitutes scientific understanding. However, we do not regard understanding merely as ”knowing” or ”believing” given propositions. Understanding is constituted by the ability to make correct what-if inferences concerning the phenomenon to be understood (Ylikoski 2009, Ylikoski & Kuorikoski 2010). The important point here is that understanding is not only about learning and memorizing true propositions, but about the capability to put one’s knowledge to use. To understand is to be able to tell what would have happened if things had been different, what would happen if certain things were changed, and what ways there are to bring about a desired change in the outcome.

Our notion of understanding is factive: it is not enough to be able to make just any inferences one wishes; one must get those inferences right. The correct what-if inferences provide a natural measure of understanding: one understands the aspects of phenomenon that one can make correct what-if inferences about, and the number and precision of correct what-if inferences determines how much one understands. This implies that understanding is not an all-or-nothing affair; rather, it comes in degrees: one can talk

about a very narrow (or even superficial) understanding, but also about deep understanding that is demonstrated by an extensive and systematic ability to answer what-if questions. Thus, one's understanding increases when the scope of the correct what-if inferences expands (Ylikoski & Kuorikoski 2010). Naturally, not all increases in understanding are equal in value: the focus should be on raising one's inferential ability with respect to those aspects of the phenomenon that are theoretically or pragmatically important.

It is an essential aspect of our account that knowledge of dependencies allows the making of counterfactual inferences. Explanatory knowledge is not about what happens regularly, but about what would happen if things were different, thus providing grounds for what-if inferences. This is the crucial difference between explanatory and merely descriptive information (Woodward 2003). This modal aspect of understanding is reflected in the contrastive-counterfactual account of explanation. Explanations, while tracking relations of dependence that exist independently of us, can only relate to things described in a certain way.[2] In other words, the target of an explanation is a specific aspect of a phenomenon, not the phenomenon as a whole. This is captured naturally by the idea that explanations are answers to contrastive questions (Garfinkel 1981, Hesslow 1983, Lipton 2004, Ylikoski 2007). We are interested in why things are one way rather than some other way; in other words, the *explanandum* is not a plain fact, but a contrastive one. The idea of a contrastive *explanandum* helps to make the explanation more explicit in an analytically fruitful manner (Ylikoski 2007). Spelling out the contrastive structure forces one to articulate what the object of the explanation is and in which respects we think the object could have been different. The fact and the foil(s) can be represented as alternative values of the same variable. Thus the foil can be an imagined (for example expected or predicted) value of the variable, or a value of the variable at some other time or in some other system. Explanation is therefore always the explanation of differences, and the task

[2] Although realist, our view is therefore not ontic in the sense of equating explanations and explanatory factors. Indeed, it would be rather strange to insist that causes and mechanisms would somehow do the explaining by themselves. As has been noted in the literature, little of substance depends on this metaphysical (or more likely grammatical) issue alone (Illari 2013).

of the *explanans* is to say what makes the difference (Mackie 1974, Woodward 2003, Waters 2007, Strevens 2008).

Explanation and understanding are not different kinds of activities, mental processes or methods. Successful explanations convey the ability to provide answers to contrastive what-if questions, and understanding is based on a body of knowledge that provides the basis for these answers. However, it is important to recognize that the inferential account of understanding does not conceive of understanding as a kind of special mental state. It is neither a privately accessible sensation or experience nor a state of mind. This is based on the fact that the criteria for attributing understanding are public. When judging whether someone understands something, people do not attempt to look into the person's mind; rather, they set out to observe whether he or she can make relevant counterfactual inferences about the phenomenon in question. Having an appropriate mental model might be a precondition of understanding, but unlike some accounts of understanding inspired by cognitive science (Waskan 2006), the inferential approach does not equate understanding with having the right kind of mental model (Ylikoski 2009; Kuorikoski 2011). After all, the *correctness* of internal mental models is judged according to manifest inferential performance, not the other way around.

Having a behavioral notion of understanding makes it possible to distinguish between actually understanding something and merely thinking that one understands it (Ylikoski 2009; Kuorikoski 2011): one of the key points of the inferential account is the distinction between understanding and the sense of understanding. The latter is the feeling that tells us when we have understood or grasped something. This sense of confidence can be easily confused with the ability that it indicates. Ideally, understanding and the sense of understanding would go hand in hand. However, empirical studies show that the sense of understanding is a highly fallible indicator of understanding. People often overestimate the detail, coherence, and depth of their understanding (Rozenblit and Keil 2002, Keil 2003). This illusion of depth of understanding can also influence scientific cognition (Ylikoski 2009) and, as we will see, it is an ever-present danger in modeling.

Another advantage of having a non-mentalistic notion of understanding is that it makes it possible to

consider cases of extended and collective understanding. Recent work on cognition has increasingly emphasized the role of external cognitive tools (extended cognition, Clark 2008) and the importance of the social division of epistemic labor (distributed cognition, Hutchins 1995) in scientific cognition. Scientific understanding is essentially collective. Scientific understanding proper is not what happens inside individual minds, but is constituted by the collective abilities of the scientific community to reason about and manipulate the objects of investigation. Given the ubiquity of extended and distributed cognition, it would be a mistake to focus the analysis exclusively on unaided individual cognition. Things that go on in individual minds are only indirectly related (through their public inferential performances) to our collective understanding of the world. The focus of this paper is on extended cognition, and we suggest that it provides a natural way to make sense of the contributions of modeling and simulation to scientific understanding.

3. Models as extended cognition

Suppose one wants to explain why cities are so often ethnically segregated (or alternatively why workplaces are segregated by gender). This is a contrastive *explanandum*, where the fact is the existence of segregation and the foil is its absence. One would like to know what makes the difference between these two states of affairs. In order to find such a variable, one must hypothesize about the possible causal processes that influence how people choose the place they live in. Such mechanistic knowledge can also be used to infer what the dependence itself would have been like, if the parts or their organization had been different. However, keeping these kinds of inferences and commitments in order is not a trivial task. The capacity of the human working memory is very limited and mental reasoning is sequential, making inferential tasks about a multitude of interdependent changes, subject to a number of constraints to be simultaneously satisfied, next to impossible to carry out by brainpower alone.

The first task is to simplify the object of analysis. While the ultimate goal is to explain segregation in a real city, like Chicago or Detroit, it is extremely difficult to gain an inferential grasp without abstracting away

from the concrete details. Cities are large, and they involve highly complex causal processes. Furthermore, there is an acute lack of information about the details of the processes. One way to proceed is to *stylize* the *explanandum* and create a simplified artificial representation of it. The idea is to only represent the crucial aspects of the phenomenon and leave everything else out. An extreme way to simplify the city is to consider it as a large checkerboard. A checkerboard is not much like a real city, but if you have two kinds of items occupying its squares, you can create observable segregation. The idea is to use this simple and easy representation as a surrogate of a real city and to study what kinds of processes can bring about segregation in it, and then hypothesize that similar processes could produce segregation in real cities.

This is the basic idea of the famous checkerboard model by Thomas Schelling[3]. His idea was to assume that the agents, which can be taken as individuals or families, populate a checkerboard-like neighborhood where some slots are empty and available for occupation by agents who are not satisfied with their current neighborhood. The agents are assumed to have preferences concerning the ethnic composition of their neighborhood and the aim of the model is to understand how a segregated neighborhood might arise from these preferences. Let us assume that there are only two kinds of agents, and let us call them A's and B's. The idea of the model is that agents will move to new (randomly assigned) slots until they are satisfied with the composition of their neighborhood.

If the agents prefer to be in the majority in their neighborhood, it is quite easy to figure out what will happen if they can freely move around. The agents' attempts to be in the majority by moving to a new location change the composition of both old and new locations. The neighborhood they leave becomes less attractive to members of their own group and the composition of the new neighborhood becomes less

[3] The checkerboard model is perhaps also the most used stock example in the philosophy of social science literature, and worries have been raised that using it repeatedly may have created biases in philosophical views. Granted, the checkerboard model might not be representative of economic and sociological models in general. However, as the model has been heralded as an example of a good explanation in social sciences (Sugden 2000, Hedström & Ylikoski 2010), it must embody at least some of the key virtues that social scientists expect their theoretical models to have. Secondly, and more importantly, in the present context it just provides a simple and well-known example to illustrate our points about using external representations in science. Nothing in our argument depends on the choice of this specific example.

attractive to the members of the other group. Intuitively it is quite clear that the agents will move around until they are satisfied in a fully segregated neighborhood. This is not a particularly exciting result, and it can be reached without a formal model. However, a more interesting question is whether there are ways to bring about segregation without people preferring it.

One hypothesis is that perhaps the simple aversion to being in a clear minority with respect to one's neighbors could be enough to lead to strongly segregated neighborhoods. A random move in an evenly distributed cityscape could then lead to someone else suddenly finding herself belonging to a local minority and thus wanting to move. This could lead to a chain reaction ending up with segregated areas, although no-one would really prefer to live in such an arrangement. The idea of a cascade like this is intuitively plausible, but is it really a possible explanation for segregation? The choices of agents are interdependent, and this makes deducing collective outcomes very difficult. Similarly, the conditions under which such cascades are to be expected (the kind of preferences that are required, how the size of the neighborhood the agents see affects the process, what kind of neighborhood structures are vulnerable, etc.) are completely beyond unaided reasoning.

A way to alleviate the cognitive load and increase the reliability of these inferences is to write down the assumptions and hypothetical changes. In this way, keeping in check what else has to be changed when a particular assumption is changed becomes manageable since the bookkeeping is externalized to an outside medium. While in principle there is nothing wrong with natural language, the language of mathematics is better suited to handling inferences regarding functional dependencies. In addition, it is also convenient in that mathematical inferences are, if done correctly, automatically truth-preserving. Using a formal language forces the reasoner to explicate implicit assumptions, and makes the comparison of inferences from similar yet different assumptions easier.

This is what many sociologists and economists following Schelling's idea have been doing (Bruch & Mare 2006, 2009; Clark & Fossett 2008; Fossett 2006; Macy & van de Rijt 2006; Muldoon, Smith & Weisberg 2012). They have shown that segregation indeed arises due to the 'tipping point' phenomenon, where the first moves of even a few dissatisfied agents can create an incentive for others to move. This results in a

cascade of relocation that only ends when the whole board has become highly segregated. One surprising finding of these models is that even if the agents only wish to avoid being in a too small minority – for example, less than a third in their neighborhood – a segregation dynamic will still emerge. Furthermore, this result is rather robust: one can make the model more complex, but the same segregated equilibrium will still follow. Thus, counterintuitively, even tolerance can lead to segregation.

Such outsourcing of inferential work to external representational aids is a paradigmatic example of extended cognition (Clark 2008). In the inferential view, it does not matter whether the model is a mental model, a physical object, a set of equations, or a computer program. What matters is the ability to use it to make correct what-if inferences about the object of interest. When the inferences from the assumptions to the *explanandum* are accomplished with the aid of an external medium (such as pen and paper, computer, or a checkerboard), and a set of formal truth-preserving inferential rules, the relevant cognitive system is not the modeler alone, but the model-modeler pair (Kuorikoski and Lehtinen 2009). This makes the extended cognitive system composed of the modeler and the external inferential apparatus the understanding subject, whose inferential performance is constitutive of understanding.

This may sound unintuitive, since we automatically associate the concept of understanding with something that takes place in the mind (or consciousness) of the scientist. As we argued above, this is a serious mistake. What is confined to the mind of the scientist is the metacognitive state of a sense of understanding. It is important to keep metacognition about understanding, and understanding itself, separate, and it is the latter that is epistemically primary. From the viewpoint of epistemic and explanatory progress, it makes little difference whether all the inferential work is confined within the brain of the scientists – and even less difference whether or not the scientists happen to experience the phenomenological sense of understanding while doing the inferential work. The goal of science is not to provide satisfying experiences to scientists – this could certainly be achieved using less resources – but to increase our collective ability to make correct what-if inferences about the world.

When the use of models is understood in terms of extended cognition, some worries about their explanatory value turn out to be misguided. For example, Alexandrova and Northcott (2013) deny that

(economic) models as such explain at all, and that they only have a heuristic explanatory use in suggesting possible explanations. In our view, models always comprise only a part of the explanatory inferential practice. Therefore the distinction between heuristic and explanatory use is vague, at best. The role of models is to facilitate inferences from assumptions, sometimes in a way which allows completely novel inferences to be made, but quite often only by forcing the modeler to be more explicit in her assumptions and making the validity of the inferences easier to check. A model, understood as an external inferential aid in an extended cognitive system, is explanatorily sterile if it does not improve the range or reliability of explanatory inferences. In such cases, the model is used mainly to illustrate the modeler's mathematical abilities and may also give rise to an illusion of depth of understanding. This might indeed be the case with many economic models, but it is not an intrinsic feature of all model-based reasoning.

Therefore there are (at least) three ways in which models can enhance scientific understanding: they oblige scientists to be more explicit about their assumptions, they can make their inferences more reliable, and they can expand the scope of correct what-if inferences that the scientists are able to make.

First, building a formal model forces theorists to be more explicit about their assumptions. In contrast to verbal theorizing, formal and computational modeling does not allow implicit assumptions within a model. Everything that goes into the formal model must be explicitly defined or coded, which, in principle, makes it easier to evaluate the assumptions from the point of view of consistency, realism, and relevance. However, it is important to recognize that model building does not completely do away with implicit assumptions. The assumptions that guide the inferences from the model to the target system are not part of the model, so they might remain implicit. For example, the model of residential segregation might not include the activities of housing agencies, while in the real world they may play an important role. Thus the (implicit or explicit) assumptions about the effects of housing agencies will guide the inferences that the theorist makes from the model to the real-world housing patterns. Formalization may also introduce further implicit assumptions that may go unnoticed especially in cases where the modeler or the modeling community does not fully understand the modeling framework itself.

Second, the use of models makes the inferences more reliable. Both psychological science and everyday

experience show that unaided human reasoning is not very reliable. We are victims to limitations in our working memory, various biases, and, of course, errors. The frequency in which the results of formal models and computer simulations are found surprising and unintuitive demonstrates the fallibility of our reasoning abilities. By externalizing some of the inferences, we can increase their reliability. Naturally, the use of external aids is not a miracle cure. For example, in the case of computer simulation, although we can almost always trust the machine to execute the code correctly, simulations have other sources of unreliability. Bugs in the code, idiosyncratic programming habits, difficulties in comparing simulations built in different programming languages (or platforms), and a lack of sufficient commentary all present challenges to simulation-aided scientific inferences.

Finally, extended cognition expands the number of what-if inferences that one is able to make, as suggested above. As also suggested above, the use of formal modeling and computer simulation removes some of the restrictions that our verbal reasoning skills place on our inferential abilities. For example, with agent-based simulations of segregation processes one can systematically study how the changes in the assumptions change the outcomes. Defining the extra understanding provided by the external aids of cognition is easy in principle: we simply compare which correct what-if inferences the scientists are able to make with and without them. Similarly, we can study whether the employment of external aids creates illusory understanding by determining whether they increase (or decrease) the number of false beliefs about our real inferential abilities.

4. Understanding the model – or the joys and follies of conceptual exploration

Models are external inferential aids used to generate explanations from the assumptions at hand. It goes without saying that using such apparatuses is not usually a trivial task. It is therefore important to distinguish the *understanding of the model* from *understanding with the model*, i.e., the empirical understanding of phenomena possibly facilitated by the model. According to the inferentialist view, the understanding of a model consists of the abilities to manipulate the external inferential apparatus: in order

to understand a model, one needs to understand how model properties change as a result of local changes in the assumptions. This involves not only the ability to "solve" the model (i.e. to deduce a specific model result), but also the ability to adapt the model (or the modeling framework) to novel situations, the ability to anticipate what kinds of ingredients fit together, and the ability to judge which parts of the model are most crucial for empirical applicability. Just as with understanding in general, understanding a model is not an all-or-nothing affair: one can understand the model to different degrees.

Usually a better understanding of the model is acquired by manipulating the model: by proving new mathematical results, by computationally exploring its properties under various parameterizations, alternative discretizations etc., and generally by simply "toying with it." Such purely non-empirical "conceptual exploration" (cf. Hausman 1992) can therefore have a perfectly legitimate epistemic role in providing a better understanding of the inferential tools, and can thus indirectly contribute to the growth of empirical understanding. Of course, such mathematical exploration can lose sight of the ultimate goal of using the models to infer about the world, and degenerate into an empirically sterile research program with little epistemic value (an accusation often raised against theoretical microeconomics).

In the case of simulation models, it is also important to recognize that simulators increasingly do not build their simulations from the ground up. Rather, they employ ready-made software libraries or modify existing simulations built by others. When this happens, it is likely that the simulator is not familiar with all of the assumptions underlying the simulation. Thus it is important to distinguish between merely using a simulation and fully understanding how it works. When a simulation is used merely as an inferential aid, the user has a very limited understanding of the underlying mechanisms. In this case, the scientist is not able to infer what would happen if some parts of the simulation were changed, nor would he or she be able to determine what the proper domain of the simulation's application is. The same point also applies to other kinds of models, although the problem is more salient in the case of computer simulations because they allow apparent inferential facility with a very shallow understanding of the model.

Understanding a model template is also often facilitated by specific narratives related to the model. Game theoretic model templates are excellent examples of this. Prisoner's dilemma, Battle of the sexes and Stag

hunt are all taught by telling characteristic little stories that lay out the essential incentive structures that define these models. Such use of narrative structures, therefore, has an important, though indirect, epistemic role in mobilizing the cognitive resources related to familiar everyday events towards the task of understanding the model (see Morgan 2012, chapters 6 and 9). Nevertheless, the acknowledgment of this indirect role does not amount to claiming that model-based explanation – understanding with the model – requires a narrative underpinning.

The achieved inferential prowess in manipulating the model itself can be mistaken for an increased understanding of the modeled phenomena – especially when accompanied by a sense of understanding. This is an important case of the illusion of depth of understanding (Ylikoski 2009; Kuorikoski 2011) related to model-based cognition. A mere increase in the proficiency of making inferences within the model does not automatically imply that the modeler also has an improved ability to make correct what-if inferences about the intended empirical target. Using the model to actually explain something also requires additional background knowledge and know-how that is not part and parcel of "the model" itself. Thus understanding of the model is only a necessary condition for understanding something with the model.

5. Understanding with the model

If the model allows drawing correct what-if inferences concerning the consequences of hypothetical changes in the modeling assumptions corresponding to explanatory factors (alternative values of variables or parameters), then it has explanatory content beyond merely a re-description and re-organization of data. The explanatory dependencies captured in a model may be causal or constitutive. But what exactly does it mean that the model "captures" such a dependency? The philosophical literature on modeling has looked for a special relation between the model and its target that could, in some sense, explain this epistemic role of the model. This relation would then be the answer to the first puzzle of modeling: how can the manipulation of these surrogate systems provide genuinely new empirical understanding of the world?

Since most models are not sentential in structure, the semantic relation of truth is not usually considered a viable candidate for this role. Indeed, the common view, also among modelers, is that all models are literally false. Since models are often taken to be mathematical structures, various mappings (isomorphism [van Fraassen 1980], homomorphism [Bartels 2006], partial isomorphism [French 2003]) have been suggested to be "the" relation explaining the epistemic role of models, but none have exhibited the right formal properties or have saved all of the central intuitions (Suárez 2003).[4] Other proposed concepts, such as similarity (Giere 2004) or credibility (Sugden 2000), are vague, and as such are mere placeholders for a proper explanatory account.

Various pragmatic approaches to representation introduce the users' intentions into the picture, thus providing the directionality of the representation relationship (Giere 2004; Mäki 2009) lacking in the structural mapping accounts. Other accounts emphasizing 'interpretation' (Contessa 2011) or 'denotation' (Hughes 1997) also depend on the intentionality of the modeler in establishing the aboutness of the model. However, if representation is grounded primarily on the specific goals and representing activities of humans as opposed to the facts about the representative vehicle and the target object, then the idea that there is something objective about the representation relation having to do with the properties of the model and the target system is missed, and it becomes questionable whether the representation relation itself does any explanatory work (Knuuttila 2011). Another danger in stressing the intentions of the modeler is conceiving representation as a mysterious mental act – as if intentionality in one's mind could magically create a relationship between a model and the world. This way of conceptualizing aboutness – conceiving intending as a special mental act that a-causally endows utterances with their meanings, or things with their representational content – was identified as a form of mythical thinking and subsequently refuted by Wittgenstein (1953). It is also worth avoiding in the context of models.

[4] Suárez (2003) presents arguments based on the variety of ways of representation, the logical properties of the representation relation, the necessity of accounting for misrepresentation, and nonsufficiency and nonnecessity arguments, to persuasively discredit proposed philosophical accounts of "the" representation relation between scientific models and their targets.

If we take model-based (explanatory) reasoning to simply be a matter of drawing conclusions from given assumptions using external inferential aids, there is nothing further to be explained about the epistemic role of models. There is no epistemic puzzle concerning how the manipulation of the surrogate system can provide new empirical knowledge about the target, since the manipulation itself constitutes an inference from the modeling assumptions to a conclusion carried out by the extended cognitive system. The only remaining epistemic questions concern the truth of the assumptions and the reliability of the inferences (Kuorikoski and Lehtinen 2009). However, if we adopt an inferential perspective on representation in particular (Suárez 2004, Vorms 2011) and aboutness in general (Brandom 1994), then the representational properties of models are accounted for in a straightforward manner: a model (as an external inferential apparatus) represents some real world phenomenon by virtue of the fact that (and to the extent that) some cognitive agent (modeler) can use the apparatus to make correct inferences concerning the phenomenon. If the new inferences made possible by the model include counterfactual what-if inferences, then the model is explanatory and represents some crucial dependencies related to the phenomenon by virtue of facilitating these inferences.

For some, explaining representation in terms of inferences seems counterintuitive or "thin" (e.g., Contessa 2011). After all, the hope was that the representation relation would explain the epistemic value of the model. Thus the philosophical deflationism of the inferentialist account of representation may feel disappointing or even problematic to some. However, to say that the relation of representation is constituted by the inferential "affordances" of the model is not to say that the relation is somehow arbitrary, "subjective," or mysterious. It is certainly not arbitrary that a specific diagram, set of equations, or physical scale model is more helpful in inferring about a specific target phenomenon than some alternatives. And this depends as much on the intrinsic properties of the inferential apparatus as it does on the cognitive make-up and perceptual abilities of the model user.[5] It is no accident that movable tokens on

[5] In this sense, the representational properties of a model are still defined in relation (and in this sense "subjective") to the cognitive agent using the model. However, there is nothing subjective or relative about the correctness, range, and reliability of the inferences carried out by the extended cognitive system.

a grid provide a good way for beings equipped with vision and hands to keep score of the rule-based movements of postulated agents, and it is no accident that such systems of reasoning can be used to make counterfactual inferences concerning segregation in cities. It is just the case that that there is no substantial and general philosophical explanation for this representational success. These (perfectly objective) dependencies between the properties of external inferential apparatuses and their possible applications, i.e., the ways in which cognitive agents can perform inferential tasks with different kinds of external aids, are empirical and therefore proper objects of study for cognitive science, not philosophy. There are genuine philosophical puzzles, but the problem of representation is not one of them.

Analyzing model-based reasoning as a type of extended cognition also provides a ready answer to the second philosophical puzzle of modeling: How can models explain, if they are always literally false because of their untrue components and explanations have to be true? Explanations have to be true in that the cited dependency must match the ontic explanatory dependence to a sufficient degree and that the derived values of the *explanandum* variable (the fact and the contrastive foil) must be "true enough." The crucial point dissolving the "paradox" (cf. Reiss 2012) is that models need not be true as a whole in order for them to yield true explanations – if they can be said to have truth values to begin with.

The main worry about models being false is not a question of imperfect accuracy as such, but of outright distortions in terms of abstractions and idealizations. Theoretical models in the social and biological sciences are almost invariably highly unrealistic, and this is not something that could be remedied by simply making more accurate measurements of key parameter values. The multi-dimensional complexity of these phenomena makes idealizations and other falsehoods necessary if the models are to be tractable or possess any degree of generality. The realism of assumptions has been a much discussed methodological issue as long as there has been modeling. Although there is some indeterminacy in the use of these terms, abstraction is usually taken to refer to the act of omitting some feature and aspect of the phenomenon, and idealization to the distortion of some aspect to a suitable boundary value (such as zero or infinity).

Idealizations are introduced into models in order to virtually/conceptually isolate the causal factors of interest from the factors which disturb or mask their manifestation in reality (Galilean idealizations), or

simply in order to render the model tractable (tractability assumptions). Galilean idealizations are explanatorily useful because they isolate a specific causal tendency from the clutter of real world phenomena (McMullin 1985). They are necessary precisely because of the essential modal element of explanatory information: the goal of explanatory models is not to provide maximally accurate and precise descriptions of occurrent events and regularities, but to trace dependencies between the selected, and hopefully important, factors.

Many of the falsities in models are there because of the requirements of mathematical tractability – the specific assumption could not be included in the mathematical model in a more realistic manner in a way that would have left the model still analytically solvable (Hindriks 2006). Models are representations built using formal languages, and this also means that assumptions have to often be implemented in the model in a more specific form than would actually be justifiable on empirical grounds. A modeled explanatory dependence is robust with respect to these tractability assumptions if the explanatory dependence is not itself dependent on the falsities in the model, i.e., the result can also be derived from the same empirically interpreted (and roughly correct) causal assumptions and alternative (but similarly literally false) tractability assumptions. Many modeling results, and even whole research programs, do not quite fulfill this desideratum. Any explanations posited on such a shaky basis should naturally be treated with caution.

The importance of robustness considerations underscores the fact that when assessing the understanding provided by theoretical modeling, the proper unit of analysis is not a single model, but rather a whole family of them (Ylikoski & Aydinonat 2014). Any explanation provided by a single model is always suspect in that the explanatory mechanism may be an artifact of false tractability assumptions. Confidence in explanations derived from idealized models should only come about through a process of robustness analysis: deriving the explanation from a set of models representing the same core explanatory mechanism but using different alternative tractability assumptions (Kuorikoski, Lehtinen, and Marchionni 2010; Levins 1966; Weisberg 2006).

The checkerboard model is very abstract and contains strong idealizations. For example, it assumes homogenous preferences, ignores strong discriminatory preferences, and incorporates very specific

assumptions concerning neighborhood structure. Even more worryingly, it completely ignores all of the well-known causes of segregation, such as economic inequalities between groups, various forms of institutional discrimination (redlining, racial steering by real-estate agents, etc.), and systematic differences in the location of employment or channels of information about available housing (Ylikoski & Aydinonat 2014). It is not surprising that these "deficiencies" have led many to dismiss the explanatory value of the model to the point of claiming that it "trivializes and endorses racism". This accusation is quite unfair, since checkerboard models have been studied for over forty years by a multitude of scholars in various disciplines and thus many of the consequent findings should not be read back to the original models. The segregation equilibrium seems to be a robust outcome in many of the modifications to the model's structural assumptions (Aydinonat 2007; Muldoon, Smith & Weisberg 2012). Most changes in the size of the neighborhood, individual preferences, or the availability of apartments do not change the segregated outcome. The robustness of the model speaks to the wide relevance of the mechanism: it can work under a number of different circumstances (Schelling 1978; Fossett 2006). However, changing the neighborhood structure, the probability of random errors, or the type of preference functions may lead to non-segregation outcomes (e.g., see Bruch and Mare 2006, 2009a; van de Rijt, Siegel, and Macy 2009). The speed of reaching the equilibrium and the kind of segregation created are much less robust attributes of these models. However, it is important to recognize that non-robust results are also interesting, and knowing more about robust and non-robust cases provides better tools for answering a broader set of what-if questions and thus improves understanding. Robustness is therefore not a strictly necessary, and certainly not a sufficient condition for a realist interpretation of a result and successful representation.[6]

What about utterly false models with no parts that could be said to represent anything? The inferential account of understanding seems to face a challenge that might be called the problem of accidental understanding. The problem arises as follows: Is it not possible for a completely wrongheaded model to

[6] More qualifications and caveats to the epistemic role of robustness analysis are provided in Kuorikoski, Lehtinen and Marchionni 2010 and 2012.

facilitate inferences that just happen to be correct by accident, and, if understanding is constituted by the ability to make counterfactual inferences, would this not mean that a completely false model could provide true understanding? But surely this would fly in the face of the idea of the factivity of explanation, and even of a realist view of science in general. The inferentialist account of representation provides a ready answer to this worry. When representational potential itself is understood as being constituted by (correct) inferential affordances, it becomes conceptually impossible for a model to be (correctly) used for explaining a real world phenomenon and at the same time fail to represent at least some aspects (the explanatory dependencies) of the target phenomenon.[7] As with understanding, the inferentialist view renders representation to be a matter of degree: models represent only some aspects of the modeled systems, and the kinds of inferences made using the model determine what these aspects are and the extent to which these inferences are correct determines how accurate the representation is. Tractability assumptions do not represent aspects of the target by virtue of the fact that the model users do not (aim to) draw direct conclusions about the target on the basis of them. If the model result changes when a tractability assumption is changed, then the result is judged to be an artifact. It is also not always clear whether a given assumption should be taken as substantive or as a tractability assumption, and this status may change over time. The line between substantial and tractability assumptions, between the representational and non-representational parts, is not fixed and is not an intrinsic property of the model, but rather a function of how the model is used in inference.

Let us summarize how the apparent air of paradox related to model-based explanation vanishes in our account. First, there is an observation: The purpose of abstract theoretical models is not to reproduce some empirical phenomenon in all its rich detail; rather, it is to capture a small set of explanatory dependencies that are assumed to be central. Second, all of the model's assumptions are not equal. For example, the truth of the model's tractability assumptions is treated differently from the assumptions that are related to

[7] There remains the logical possibility that a model originally not intended to represent a specific system or one built using unsound epistemic principles could reliably facilitate correct counterfactual inferences by purely accident. We do not know of any cases of such accidental representational success,

the explanatory dependence the model is trying to capture. The point of robustness analysis is to show that tractability (and other non-essential) assumptions make no difference to the explanatory dependence. This does not obviate the assumption that the model should get the target explanatory dependence right. Thus explanation remains factive. Finally, when an abstract theoretical model is used to explain some concrete empirical phenomenon, the claim is not that the model provides a full or complete explanation of a puzzling fact. Rather, the suggestion is much more modest: the model captures an important, maybe even crucial, element of the phenomenon's explanation. It is a bet on the claim that the mechanism captured by the model plays a central role in the full explanatory story. Naturally, this bet can be misplaced as the mechanism may not have the suggested role. In such cases the model provides at most a part of a how-possibly explanation (Ylikoski & Aydinonat 2014), not the actual explanation of the phenomenon. Thus, when understood correctly, theoretical models provide no special explanatory challenges or paradoxes for the theory of explanation.

6. Evaluating explanatory models

We have argued above that our inferential view can answer the most vexing philosophical problems related to model based explanation. If possible, a theory of explanation and understanding should not only tell when and in virtue of what something is or is not an explanation, but also provide tools to assess when one explanation is better than another. We now show how the view can be used to comparatively evaluate the explanatory qualities of models. The explanatory potential of a particular inferential apparatus can be evaluated along the lines laid out at the beginning of this paper: the more (and the more interesting) what-if questions the model helps the model user to answer, the more understanding it provides. This is an immediate consequence of our inferentialist view, since understanding is *constituted* by these inferential abilities. This increase in explanatory inferential ability can be broken down along different *dimensions of explanatory power*: non-sensitivity, precision, factual accuracy, degree of integration, and cognitive salience (for a more comprehensive treatment, see Ylikoski and Kuorikoski 2010). The first four dimensions mainly concern the number and quality of the inferences made possible by the model, whereas the last one

is about the ease and reliability of drawing these inferences. The dimensions are partly independent, and there are systematic (though non-strict) trade-offs between them. The trade-offs can also be found in the case of model-based explanations. Here we discuss only trade-offs among explanatory virtues and do not take a stand with respect to the much discussed issue of strict trade-offs among general model desiderata inspired by the work of Richard Levins (see, e.g. Matthewson and Weisberg 2009).

Formal models are more exact than purely verbal theorizing in that the logical and mathematical relationships between concepts and variables are explicitly defined, thus making it possible to derive more precise implications from the assumptions. However, these same models might not fare well with respect to facilitating *detailed* inferences about the *explanandum* phenomenon. Thus they are often deficient in the dimension of explanatory power that we call *precision*. The key here is to understand that the exactness of formal and mathematical derivations is not the same property as the dimension of explanatory goodness that we have labeled precision. Theoretical models can rarely credibly aim at providing quantitative estimates about the importance of the modeled factors. For example, the segregation models can only provide grounds for qualitative reasoning that some factors are relevant to, say, the speed or probability of spontaneous segregation, but not to what degree. In other words, while the model is formally exact, it is not very precise in being able to account for fine grained contrasts in the empirical *explanandum* phenomenon. The model says that segregation is to be expected in a wide range of circumstances, but the more precise characteristics of the produced segregation will remain opaque. This is a quite general property of abstract theoretical models. Outside physics, with its stable, universal, and precise physical constants, any quantitative inferences must be made on the basis of case-specific quantitative data and this is not possible with models that ignore many details of the particular setting in order to capture something more general about the mechanism of interest.

This is the reason why theoretical models, being highly abstract, would usually not be considered very interesting if they were highly sensitive to specific initial conditions – i.e., highly *factually accurate*. A model must be robust with respect to (at least some) of the idealizations that are known to be always false, but it can also be robust with respect to assumptions that can be true or false, depending on the empirical

application. In practice, these kinds of robustness are not always easy to distinguish between. Furthermore, as stated above, the interpretation of the assumptions might change: some assumptions that were originally treated as pure tractability assumptions may later be interpreted as empirical assumptions, or vice versa. Nevertheless, the more robust the result, the better the explanations derived from the model. Note that robustness is important for two reasons. First, it is needed to justify the extrapolation from an abstract model to a specific application – if the result is not robust there is no reason to expect that it would hold in situations where idealizations and omissions of the model do not apply. This is a purely epistemic requirement related to how reliably the model result can be used in explaining real world phenomena, not an explanatory virtue per se. Second, robustness is also related to *non-sensitivity* as a dimension of explanatory goodness: *given that* the explanation is true, the more sensitive the explanatory dependence is to changes in background factors, the less powerful explanations it provides. In either case there is a trade-off: robustness can usually only be gained by sacrificing precision and factual accuracy (Kuorikoski, Lehtinen & Marchionni 2010).

Some models are built anew from the ground up while others belong to a research program using a family of established templates. If the model is built from scratch, the modeler, provided she is clever enough, can model precisely those and only those factors she hypothesizes to be important. The downside is that relating such stand-alone models, especially theoretical ones, to other bodies of model-based knowledge is often difficult. If there is only one isolated model, it is next to impossible to evaluate the importance of the factors modeled and the reliability of the inferences given the idealizations implemented in the model. This is one way in which models differ in their *degree of integration*. If a model is integrated within a program of similar models, then the reliability of the explanatory inferences becomes easier to evaluate. The extended model family can also provide additional "theorems" and suggestions about how to expand the range of what-if inferences derivable using the model. The continuing research on Schelling type models has revealed that although segregation is a robust result under many structural assumptions, changes in neighborhood structure, the probability of random errors, or the type of preference functions may lead to non-segregation outcomes – and some of these dependencies may have a realistic interpretation (e.g.,

Bruch and Mare 2006; 2009). Thus an integrated family of models contributes to explanatory knowledge more than individual models considered separately.

These considerations make it understandable why scientists do not feel that developing hyper-realistic models is a promising way to gain explanatory insight into complex systems. While replacing idealizations with empirically established values and adding omitted factors might make the model a more realistic and truthful representation of the observable attributes of a target system, such a model might not improve explanatory understanding. Apart from the fact that scientists are building the model partly because they do not know which omitted factors are relevant, this complex model would also be deficient in the understanding it provides. There are two major reasons for this. First, a complex model would be more difficult to handle, in other words, the scientists' ability to perform what-if inferences of interest might be seriously reduced. Understanding with the model usually remains shallow and unreliable if it is not accompanied with a proper understanding of the model. Second, the aim of explanatory models is to capture the key dependencies on which the behavior of the system depends, but a hyper-realistic model that manages to "save the observable phenomenon" might just obscure them. Thus one would not achieve a general understanding of the underlying causal mechanisms that could also be applicable to other cases. This is important because the aim of an explanatory model is to represent explanatory dependencies relevant to the phenomenon, and not simply represent one particular instance of the phenomenon. Thus one must be selective, and some idealizations and omissions may in fact be virtuous from the point of view of representing explanatory relations.

Models differ in their ease of use, in the efficiency of the interface between the model and the model user. This *cognitive salience* of models is of direct epistemic relevance, since it affects the number and reliability of inferences available to the extended cognitive system. Cognitive salience can be based on anything from a good user interface in a simulation package to a particular way of representing a mathematical structure that makes the identification of the necessary empirical background assumptions as transparent as possible. Naturally, the ease of use also depends on the background skills of the scientist: her judgment of the relative explanatory merits of a specific model may be based on her familiarity with the modeling

framework – she is able to think with it – rather than the more general deficiencies in competing models. It might be that in the hands of other scientists, the competing models may provide comparable explanatory insight.

Sometimes the ease of use increases the chances of the illusion of depth of understanding by making "toying with the model" too easy in such a way which crowds out thinking about the crucial background assumptions needed to reliably infer with the model to a real world phenomenon. As mentioned above, off-the-shelf micro-simulation packages in the social sciences are a good example of this. In contrast, programming the simulation from the ground up necessitates thinking about what exactly the important assumptions are, what kind of empirical background assumptions have to be made in order for the model to be interpretable, and what kind of tractability assumptions have to go into the model to make it computable.

The cognitive salience and degree of integration also (partially) account for the widely held view that using a single unifying model template to explain many phenomena is explanatorily virtuous. If the same model template can be used in a multitude of situations to make novel inferences about different phenomena, then obviously the model user is spared from the work of learning to use new inferential aids. Using a similar model also facilitates linking related results and inferences in the analogous domains of application, thus possibly enabling new counterfactual inferences and increasing the degree of integration. Nevertheless, this does not imply that unification as such is constitutive of explanatory understanding, as this easy applicability of the model might be achieved by a serious simplification or misrepresentation of the original *explanandum* phenomenon. In such cases, increased "unification" does not increase understanding in the sense of the ability to make correct what-if inferences about the phenomenon. This suggests that philosophical theories that attempt to reconstruct explanatory understanding in terms of unification might be based on an attribution error. Moreover, it is important to keep in mind that such unification may come at a great cost with respect to the other dimensions of explanatory goodness, and can easily lead to illusory understanding. For example, after becoming proficient in using the segregation simulations, our urban sociologist might begin to see segregation processes everywhere. This is not a bad

thing in itself, since segregation processes in the abstract sense are everywhere. However, the habit of explaining diverse phenomena with the same segregation model becomes problematic when a specific mechanism for segregation is automatically taken to be the best or even the only explanation for segregation outcomes just because that would be "unifying" or "elegant."

Conclusions

So how do unrealistic surrogate systems explain real phenomena? Above we have laid out an inferentialist answer to this question. A model can be said to explain a contrastive *explanandum* if it can be reliably used as an inferential aid in correctly reasoning that if the *explanans* variable had been different, the *explanandum* would have been different as well. Such an inferential apparatus is explanatorily valuable if it increases the range of such explanatory inferences, makes the inferences more reliable or helps to explicate the conditions under which such inferences can be made. The explanatory power of the model, and consequently the amount and type of understanding it can provide, amounts to the number and importance of these inferences it enables. This explanatory power can be further analyzed according to our schema of the dimensions of explanatory power. Our inferentialist framework also captures the epistemic role of understanding the model: the range and reliability of the explanatory inferences made possible by a representational artifact depends on the interface between the user and the model. Understanding the model and understanding with the model should be kept separate, since the intuitive sense of understanding triggered by understanding the model may be confused with an increased understanding of the modeled phenomena.

The "representational power" of the model is also constituted by its inferential potential, not the other way around. This is not to say that the representation relation is somehow arbitrary or depends on the whim of the model user – it is just that there is nothing philosophically substantial to say about the ways in which agents with particular cognitive abilities can use external tools to extend their inferential and manipulative capabilities concerning their surroundings as well as themselves. Conceptualizing modeling in terms of

extended cognition and inferential activity helps to keep intuitions about the analogy between modeling and experimentation in check. The epistemic (inductive) gap is not really between the manipulated artificial world of the model and the real world, but between the set of modeling assumptions, which always includes literally untrue and even empirically uninterpretable members, and the conclusions. Models are arguments, not experiments.

Acknowledgements

This paper is based on a presentation given in a Workshop on Explanatory Power at Ruhr University Bochum in 2012, and presented in the Philosophy of Science Seminar at the University of Helsinki in 2013. We thank the audiences of these events, as well as the reviewers, for their valuable comments.

References

Aydinonat N. E. (2007). Models, conjectures and exploration: An analysis of Schelling's checkerboard model of residential segregation. *Journal of Economic Methodology*, 14, 429–454.

Alexandrova, A. and R. Northcott (2013). It's just a feeling: why economic models do not explain. *Journal of Economic Methodology,* 20, 262-267.

Bartels, A. 2006. Defending the Structural Concept of Representation, *Theoría*, 21, 7-19.

Brandom, R. 1994, *Making it Explicit: Reasoning, Representation and Discursive Commitment*, Cambridge MA and London: Harvard University Press.

Bruch, E. & R. Mare 2006, Neighborhood choice and neighborhood change. *American Journal of Sociology*, 112, 667–709.

Bruch, E. & R. Mare 2009, Preferences and pathways to segregation: Reply to van de Rijt,Siegel, and Macy. *American Journal of Sociology*, 114, 1181–1198.

Clark, A. 2008, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press.

Clark, W. & M. Fossett 2008, Understanding the social context of the Schelling segregation model. *PNAS*, 105, 4109–4114.

Contessa, G. 2011, Scientific Models and Representation. In Steven French & Juha Saatsi (eds.), *The Continuum Companion to the Philosophy of Science*. Continuum Press.

Fossett, M. 2006, Ethnic preferences, social distance dynamics, and residential segregation: Theoretical explorations using simulation analysis. *Journal of Mathematical Sociology*, 30, 185–274.

French, S. 2003, A model-theoretic account of representation (or, I don't know much about art … but I know it involves isomorphism), *Philosophy of Science,* 70, 1472–1483.

Garfinkel, A. 1981, *Forms of Explanation: Rethinking the Questions in Social Theory*, Yale University Press.

Giere, R. 2004, How Models Are Used to Represent Reality, *Philosophy of Science,* 71, S742-752.

Hausman, D. 1992, *The Inexact and Separate Science of Economics*. Cambridge University Press.

Hedström, P. & P. Ylikoski 2010, Causal mechanisms in the social sciences. *Annual Review of Sociology*, 36, 49–67.

Hesslow, G. 1983, Explaining Differences and Weighting Causes. *Theoria* 49 (2):87-111.

Hindriks, F. A. 2006, Tractability Assumptions and the Musgrave–Mäki Typology, *Journal of Economic Methodology*, 13, 401–23.

Hughes, R. 1997, Models and Representation. *Philosophy of Science,* 64 (Supplement), S325-S336.

Hutchins, E. 1995, *Cognition in the Wild*, The MIT Press.

Illari, P. 2013, Mechanistic Explanation: Integrating the Ontic and Epistemic, *Erkenntnis,* 78: 237-255.

Keil, F. 2003, Folkscience: coarse interpretations of a complex reality, *Trends in Cognitive Sciences,* 7 (8), 368-373.

Knuuttila, T. 2011, Modelling and Representing: An Artefactual Approach to Model-based Representation. *Studies in the History and Philosophy of Science* 42: 262-271.

Kuorikoski, J. 2011, Simulation and the Sense of Understanding. In Paul Humphreys & Cyrille Imbert (eds.), *Models, Simulations, and Representations*. Routledge: 168-187.

Kuorikoski, J. & A. Lehtinen 2009, Incredible worlds, credible results. *Erkenntnis*, 70, 119–131.

Kuorikoski, J., A. Lehtinen & C. Marchionni 2010, Economic modeling as robustness analysis. *The British Journal for the Philosophy of Science*, 61, 541–567.

Levins, R. 1966, The Strategy of Model Building in Population Biology, *American Scientist*, 54, 421–31.

Lipton, P. 2004, *Inference to the best explanation*. Abington: Routledge.

Mackie, J.L. 1974: *The Cement of the Universe*. Oxford: Oxford University Press.

Macy, M. & A. van de Rijt 2006, Ethnic preferences and residential segregation: Theoretical explorations beyond detroit. *Journal of Mathematical Sociology*, 30, 275–288.

Matthewson and Weisberg, 2009, The structure of tradeoffs in model building, *Synthese,* 170 (1), 169-190.

McMullin, E. 1985, Galilean Idealization. *Studies in History and Philosophy of Science Part A,* 16, 247–73.

Morgan, M. 2012, *The World in the Model: How Economists Work and Think*. Cambridge and New York: Cambridge UP.

Muldoon, Smith & Weisberg, 2012, Segregation that no one seeks. *Philosophy of Science*, 79, 38–62.

Mäki, U. 2009, MISSing the world: Models as isolations and credible surrogate systems. *Erkenntnis*, 70, 29–43.

Reiss, J. 2012. The Explanation Paradox. *Journal of Economic Methodology,* 19, 43-62.

Rozenblit, L. and F. Keil 2002, The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science,* 26, 521–562

Schaffer, J. 2005, Contrastive Causation. *Philosophical Review,* 114 (3), 327-358.

Schelling, T. 1978. *Micromotives and macrobehavior*. London & New York, NY: W. W. Norton.

Strevens, M. 2008, *Depth: An Account of Scientific Explanation*. Harvard: Harvard UP.

Suárez, M. 2003, "Scientific representation: Against similarity and isomorphism", *International Studies in the Philosophy of Science,* 17, 225–244.

Suárez, M. 2004, An Inferential Conception of Scientific Representation. *Philosophy of Science,* 71, 767-779.

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

van Fraassen, B. 1980, *The Scientific Image*. Oxford: Oxford University Press

Vorms, M. 2011. Representing with Imaginary Models: Formats Matter. *Studies in History and Philosophy of Science* 42:287-295.

Waskan, J., 2006. *Models and Cognition: Prediction and Explanation in Everyday Life and in Science. Cambridge*, MA: The MIT Press

Waters, K. 2007, Causes That Make a Difference. *Journal of Philosophy,* 104 (11), 551-579

Weisberg, M. 2006, Robustness Analysis, *Philosophy of Science*, 73, 730–42.

Woodward, J. 2003, *Making things happen: A theory of causal explanation*. Oxford and New York: Oxford University Press.

Ylikoski, P. 2007, The Idea of Contrastive *Explanandum*. In Johannes Persson & Petri Ylikoski (eds.), *Rethinking Explanation*. Springer.

Ylikoski, P. 2009, The illusion of depth of understanding in science. In H. De Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding: Philosophical perspectives* (pp. 100–119). Pittsburgh: Pittsburgh University Press.

Ylikoski, P. 2013, Causal and constitutive explanation compared. *Erkenntnis,* 78, 277-297.

Ylikoski, P. 2014, Agent-based simulation and sociological understanding. *Perspectives on Science,* 22, 318-335.

Ylikoski, P. & Aydinonat, E. 2014. Understanding with Theoretical Models. *Journal of Economic Methodology* 21: 19-36.

Ylikoski, P. & J. Kuorikoski 2010, Dissecting explanatory power. *Philosophical Studies*, 148, 201–219.