sciendo

What is the Problem with Model-based Explanation in Economics?

Caterina Marchionni
University of Helsinki

BIBLID [0873-626X (2017) 47; pp. 603–630]
DOI: 10.1515/disp-2017-0020

Abstract
The question of whether the idealized models of theoretical economics are explanatory has been the subject of intense philosophical debate. It is sometimes presupposed that either a model provides the actual explanation or it does not provide an explanation at all. Yet, two sets of issues are relevant to the evaluation of model-based explanation: what conditions should a model satisfy in order to count as explanatory and does the model satisfy those conditions. My aim in this paper is to unpack this distinction and show that separating the first set of issues from the second is crucial to an accurate diagnosis of the distinctive challenges that economic models pose. Along the way I sketch a view of model-based explanation in economics that focuses on the role that non-empirical and empirical strategies play in increasing confidence in the adequacy of a given model-based explanation.

Keywords
Economic models, explanation, idealizations, Schelling's model, Prisoner's Dilemma

1 Introduction

There has been much debate recently in the philosophy of economics about how idealized models can be explanatory. In addition to the obvious opposition between those who hold that such models can be explanatory and those who argue against it, typically pointing to the ineliminable presence of false assumptions, others have sought to develop accounts positing that idealized models could enhance understanding of phenomena without explaining them. Much of this debate has proceeded on the assumption that, given the obvious re-

Disputatio, Vol. IX, No. 47, December 2017
Received: 05/09/2017 Accepted: 02/11/2017

quirement of explanation that the explanans is true, either a model's assumptions are true and therefore it provides the actual explanation, or its assumptions are false and hence it cannot explain. Reiss (2012) claims, for example, that economic models pose a conundrum for explanation: in that they contain false assumptions they cannot be said to be true, yet only true accounts are explanatory. Along the same lines, Alexandrova and Northcott (2013: 262) note that models "do not qualify as causal explanations because they are false and therefore they do not identify any actual causes" (see also Grüne-Yanoff 2013: 257, Rice 2015: 589). Similarly, Reutlinger et al. (2017) write:

> …generally, toy models do not satisfy the veridicality condition. For instance, the DY [viz. Dragulescu-Yakovenko] model is idealized, as it assumes that economic agents are all identical, have no expectations and "zero intelligence". This is certainly an assumption that we deem (and surely hope) to be literally false. Hence, it is, at least, questionable whether the veridicality condition is met[1] (Reutlinger et al. 2017: 15).

Framing the problem in terms of falsity of the assumptions and hence of falsity of the explanation, as these authors seem to do, obfuscates the distinction between two separate issues that are both relevant to an assessment of a model's explanatory power: the conceptual issue concerning what must be the case for the model to be explanatory, and the epistemological issue concerning the justification for believing that the model possibly, probably, or very probably satisfies those requirements (see also Mäki 2013). Whereas explanantia are either true or false, and hence a model either provides or does not provide an explanation, only in a few cases do we know whether the explanantia in the model represent the actual causes that make a difference to the explanandum. Treating both the conceptual and the epistemological issues as dichotomies has the unfortunate effect of lumping together different problems that modeling in economics might have: one is that by their very nature idealized models cannot provide explanations; another is the extent to which the practice of economic modeling keeps theoretical models shielded from empirical evidence; and yet another is whether the casual empiricism that

[1] The DY model treats economic exchanges as analogous to molecular collisions in a gas and applies the tools of statistical mechanics to derive features of the distribution of individual monetary incomes (Reutlinger et al. 2017).

seems to characterize economics is sufficient to justify a belief in the explanations economic models provide.

My focus here is on theoretical models. Such models are typically abstract—they ignore many details about their target—and idealized—they misrepresent features of their target, typically to neutralize the effect of factors that are deemed explanatorily irrelevant. Unlike empirical models, which test theories or measure and estimate relationships between variables by using empirical data, theoretical models typically aim at investigating the qualitative consequences that derive from a given set of theoretical assumptions, such as about agents' behavior and their interactions. Although the principles of economic theory often guide the construction of such models, these are generally insufficient to derive any interesting result. Hence, theoretical models in economics tend to include a host of *tractability assumptions*, namely assumptions introduced solely for the purposes of mathematical tractability (Hindriks 2006).

In this paper I do not aim to offer a novel account of how theoretical models explain. On this I rely on the already existing view according to which models explain when they successfully represent some of the causes of, or the mechanism responsible for bringing about the target phenomenon (see e.g. Hausman 2013, Mäki 2013). My goal is more modest. It is to clarify the terms of the philosophical debate concerning whether economic models can explain, and thereby to make progress in understanding the problems with model-based explanation in economics: that is, whether the problem is that economic models in general cannot provide actual explanations, and hence that their contribution must lie elsewhere, or whether the problem is that some peculiarities of economics makes it particularly difficult to assess whether or not its models explain. I will suggest that the latter is often the case.

The rest of the paper is organized as follows. Section 2 reviews three peculiarities of the practice of economic modeling that seem to make the challenge of explanation more pressing. Section 3 tackles the conceptual question of what conditions a model must satisfy to be explanatory. A number of distinctions relevant to identify the explanatory power of models are introduced. Section 4 employs these distinctions to examine debates about model-based explanation in two cases: Schelling's model of spatial segregation and Axelrod's

explanation of WW1 truces with reference to the Prisoner's Dilemma. Section 5 examines the epistemological question in more detail. First, it describes the epistemic benefits of derivational robustness analysis (the practice of probing the robustness of results to the model's assumptions) then it brings the issue of empirical support to bear on model-based explanations. Section 6 reconsiders Schelling's model and Axelrod's explanation from the epistemic perspective offered in the previous section. Section 7 concludes the paper.

2 Three peculiarities of theoretical modeling in economics

The question of whether highly idealized models can explain real-world phenomena arises from the tension between the presence of idealizations that introduce elements of falsehoods and the fact that, on many accounts of explanation, genuine explanation requires the truth of its explanantia. Notice that this challenge is not limited to economics. Some characteristics of economics appear to make some of the standard ways of addressing this challenge inapplicable, however (see also Basso, Lisciandra and Marchionni 2017). First, in economics it is seldom possible to de-idealize and hence make the models progressively more realistic. Moreover, economic theory does not indicate how the idealizations in the models will affect their results. According to Nancy Cartwright (2009: 45), "In the case of economic models it is clear by inspection that the unrealistic structural assumptions of the model are intensely relevant to the conclusion. Any inductive leap to a real situation seems a bad bet." Why should we believe that the model result applies to a real-world target when we know it crucially depends on assumptions known to be false? Hence, the first peculiarity of economics is that not only do economic models, like all scientific models, contain false assumptions, but such *assumptions are indispensable for the derivation of the results*.

The second peculiarity of economics is the *tenuous relation between theoretical models and empirical models*. Here is how Roger Backhouse summarizes the problem:

> In an ideal world, the models that economists confront with data would be the same as their theoretical models. In practice this is not always possible: theories may involve unobservable variables; other

> variables may not be measurable or measured properly, theories may specify functional forms that cannot be estimated given the available techniques, and theories may simply be too complicated or too imprecise to be testable. The result is that, probably in most cases, the model that is tested is not the same as the one that is produced by the theory. It may not even be a special case of the theoretical model, but one that has been modified in ways that make it possible to confront it with data (Backhouse 2007: 145).

The tenuous relation between theoretical models and the empirical models that are eventually confronted with the data implies that only rarely is it possible to test the theoretical models for their fit with their real-world targets (cf. also Cartwright 2002).

The third feature that has been noticed to be distinctive of economic modeling, namely *its casual empiricism,* is partly a consequence of the second one. Backhouse (2007) suggests that theoretical models are often only tested "informally". Economist Dani Rodrik (2015) also points to the informal ways in which empirical evidence is often brought to bear on the models of economics. According to Anna Alexandrova and Robert Northcott, this should be regarded as a problem.

> As a general matter, the economics profession is known for its 'casual empiricism.' As the name suggests, it involves scoring explanatory victories casually rather than relying on econometric or experimental tests. Often this involves nothing more than drawing a vague and intuitively appealing analogy between the model and the phenomenon (Alexandrova and Northcott 2013: 264).

Referring specifically to theoretical models, Till Grüne-Yanoff (2009: 88) comments thus: "Economic modelers often do not refer either to data or other established and particular real-world facts, or to established regularities about sets of real-world phenomena when constructing and presenting their models." Hence, the problem is not only that it is hard to test theoretical economic models directly, but also that economists tend to be content with casual empiricism, and sometimes in more extreme cases do not even mention any real-world phenomena.

Theoretical modeling in economics thus seems to raise somewhat different concerns related to explanation: whether the models can be explanatory despite the presence of false assumptions; whether

they can be explanatory despite the presence of false assumptions known to be relevant to the conclusions; whether in the absence of direct empirical confirmation their explanations are to be trusted; and whether casual empiricism is sufficient to justify a belief in the explanations they provide. My contention is that clearly separating conceptual questions about what conditions a model should satisfy in order to count as explanatory and epistemological questions about what sort of evidence can be used to determine whether the model satisfies those conditions and hence it is explanatory helps achieving a more accurate diagnosis of the problems with model-based explanation in economics.[2]

3 The structure of model-based explanations

In this section, I deal with the conceptual question of what conditions a model should satisfy in order to be explanatory. The truth of the explanantia is one of the conditions that an account has to satisfy in order to count as an actual explanation, although the presence of falsehoods in a model does not necessarily prevent it from being explanatory. There are several ways in which a model can be said to be explanatory, which are compatible with the presence of at least some falsehood. Each of these ways in turn requires different success conditions and poses different challenges for finding out whether those conditions are satisfied.

For the purposes of this paper, I take explanation to be a matter of citing the factors that make a difference to their effects. Models partake in explanation by showing how changes in these factors make such a difference (e.g. Woodward 2003). In keeping with this account, the challenge posed by idealized models is as follows: how can models with assumptions that are false about their targets capture actual difference makers? Addressing this question requires giving an account of what it means for a model to be explanatory, which I offer in this section, and one that indicates what sort of evidence can be used to determine whether the model is probably explanatory, which I outline in Section 5.

[2] These formulations are inspired by Kirkham (1992)'s distinction between semantic and epistemic questions about truth.

What does it mean for a model to be explanatory? At one extreme one could hold that a model explains only when the model is *identical* to the actual explanation. This view is unnecessarily demanding, however: any false assumption in the model would imply the falsity of the explanans (cf. Rohwer and Rice 2016, see also Hausman 2013, Mäki 2013). However, some idealizing assumptions are made for the very purpose of leaving out factors that are deemed to be explanatorily irrelevant. Hence, the presence of at least some kinds of falsity should be allowed. At the other extreme, one could hold that a model is explanatory if in one way or another it contributes to building real-world explanations. This seems too liberal, however, as it makes sense to talk of model-based explanation only when the model is indispensable to the explanation.

In between these extremes lies the view that a *model is explanatory when it provides the actual explanation* and it does so in virtue of successfully representing the causes that make a difference to the target phenomenon. The advantage of this view is that there need not be a one-to-one relation between properties of models and properties of explanation, and therefore we can discard at the outset the thesis, probably endorsed by no one, that the presence of any false assumption automatically entails the falsity of the explanation. The problem with this formulation is that requiring a model to provide *the* actual explanation of the phenomenon remains ambiguous. It may mean that the model should identify all the causes of a phenomenon, the most important causes, or some of the causes. Alternatively, it may require it to provide the best explanation of the phenomenon, or to have good enough reason to believe in the model-based explanation. To resolve this ambiguity, it is helpful to distinguish different attributes a scientific explanation, whether model-based or not, can have.

An explanation comprises an explanans and an explanandum. A *complete* explanation includes all and only the causes that do make a difference to the explanandum, whereas a *partial* explanation includes only one or some of the causes that make a difference. An explanation is *potential* rather than *actual* when it is not known whether the explanans satisfies the truth condition (Hempel 1965).[3] An ex-

[3] Hempel (1965) formulated these requirements for the D-N model of explanation, but they have been reformulated to apply to causal explanations. It is

planation can then be both *potential* and *partial* at the same time (see also Aydinonat 2008). The second component is the explanandum. Explananda may be specific instances of a phenomenon such as some aspect of the actual pattern of ethnic segregation in Philadelphia, or generic phenomena such as some aspect of spatial segregation. The latter are the kind of explananda with which theoretical economics is often concerned: they are the stylized facts economists aim to explain. These two types of explananda parallel Weisberg's (2013) distinction between *target-directed modeling* and *generalized modeling*. The latter targets the features that are either shared or similar across various specific systems, and is the common modeling strategy in theoretical economics. A given explanation can display any combination of these properties: potential or actual; complete or partial; specific or generic.

Furthermore, explanations can differ in terms of the kind of information they provide and of the degree of explanatory power they enjoy. First, causes can be described at various levels of abstraction and detail, meaning that the level of description at which the causes are described determines the kind of explanatory information that is provided. The idea that there are different kinds of explanatory information is not new (e.g. Garfinkel 1981, Jackson and Pettit 1990, Sober 1999, Marchionni 2008). Proposals differ on the specifics, but what they have in common is the recognition that abstraction from the details of a specific occurrence to focus on general features is explanatorily valuable in itself. Jackson and Pettit (1990), for example, distinguish between explanations that provide *fine-grained information*, in other words causal information about the specific features of a specific occurrence of the phenomenon, and those that provide *coarse-grained information*, in other words causal information about what features a range of different occurrences have in common

---

also rather straightforward to reconsider the requirement of completeness to take account of the pragmatic dimension of explanation. Pragmatic considerations can be built into the explanandum, for example, by specifying both the aspect of the phenomenon in need of explanation as well as a contrast class. The completeness condition is meant to highlight the fact that an explanation might only focus on a small subset of what makes a difference to the explanandum thus specified. Hence, even within a pragmatically delimited context there is a sense in which one can talk of more-or-less complete explanations.

(Jackson and Pettit 1990). Coarse-grained causal information is obtained by way of abstracting from the specificities of particular instances of the phenomenon. Therefore, there are at least two complementary ways of explaining a particular instance of the phenomenon (for example, the actual pattern of ethnic segregation in Philadelphia or the occurrence of truces in World War One trenches).

Second, explanatory power comes in degrees. Two accounts of the same phenomenon might both be explanatory, but one might provide a better explanation than the other along some dimension. Criteria of explanatory power vary depending on which theory of causal explanation is subscribed to. For the sake of simplicity, it is sufficient to allow that some explanations might be better because they are more detailed, they rely on more robust generalizations, or they are more complete.[4] Thus, explanations that are very sparse in terms of details might be deemed less powerful than those that include more relevant details. Notice that one could also claim that one explanation is better than another when it is more strongly confirmed. Nevertheless, it is useful to keep these two senses of explanatory power separate: two compatible explanations might both be true, yet one is deemed better than the other because it is more detailed (Ylikoski and Kuorikoski 2010).

These distinctions indicate that there are different ways in which a model can be explanatory: it could provide the complete or a partial explanation, offer fine-grained or coarse-grained explanatory information, or give a more-or-less detailed explanation. As a way of capturing the different features an explanation can have, we can relax the requirement for a model to be explanatory as follows: a model is explanatory when it provides explanatorily relevant information in virtue of successfully representing some of the causes that make a difference to the explanandum phenomenon. Clearly, that the causes the model identifies are actual difference makers is one of the conditions a model has to satisfy in order to be explanatory.

[4] Ylikokski and Kuorikoski (2010) identify multiple dimensions of explanatory power. Marchionni (2013) applies the idea that explanatory power has multiple dimensions to the modeling of networks. Northcott (2013) offers a formal treatment of the notion that explanations may have different degrees of partiality. The observation that some explanations might be more powerful than others along some dimension, such as level of detail, is sufficient for the present purposes.

Yet, depending on the kind of explanation, different requirements on similarity between the model and the target have to be satisfied.

4 Two examples of model-based potential explanations

Equipped with the conceptual tools laid out in the previous section I now examine two well-known models, namely Schelling's model of segregation and the Prisoner's Dilemma, which have been claimed not to be explanatory in spite of appearances to the contrary. My aim is to show that two different kinds of arguments have been advanced to cast doubt on the models' explanatory power: for some, this kind of models cannot provide *actual* explanations, for others what is missing is the justification for believing that they do. In this section I will mainly focus on clarifying the kinds of explanation the two models can be taken to provide, and in Section 5 I will come back to the question of whether it is legitimate to believe that they succeed in picking out actual difference makers.

4.1 Schelling's model of segregation

Schelling's (1978) model of segregation starts from a random distribution of agents in a checkerboard-like city. The agents are of two colors, black and white, and may either stay put or move at no cost. If they have a preference for not being in a minority in their neighborhood, and change neighborhoods when their preference is not satisfied, it is easily shown that a pattern of segregation emerges and is stable. The ground-breaking insight of Schelling's model is to demonstrate that neither ethnic nor other forms of spatial segregation need emerge from strong discriminatory preferences. Although it is known that segregation is often the result of discriminatory preferences, preferences about the ethnic composition of a neighborhood are not the only determinants of where one chooses to live, and cities do not look like checkerboards. Considerations such as these have led some commentators to claim that Schelling's model provides a how-possibly explanation rather than a how-actually explanation (e.g. Grüne-Yanoff 2009, 2013; Ylikoski and Aydinonat 2014; Weisberg 2013; Reutlinger et al. 2017).

According to Grüne-Yanoff (2013), for example, rather than

explaining segregation, Schelling's model produces a change in confidence in the belief that racist preferences are among its necessary causes. The reason is that Schelling's model "is not established as an adequate representation of any real-world system" (855). The only component that Schelling explicitly links to the world is the assumption that individuals have preferences for not being in a minority, but no attempt is made to connect either the migration process or the checkerboard to a real-world target. The claim that Schelling's model is not established as an adequate representation, and hence cannot be regarded as providing a how-actually explanation, could be interpreted in a number of ways. First, Schelling's model has no representational relationship with a real-world target, and therefore does not satisfy one of the requirements set for a model to provide a possible explanation. Second, the model fails to represent its target, and hence we are not justified to conclude that it explains segregation. Third, the representational relationship has not yet been established, and hence we do not know whether the model is explanatory or not.

The problem with the first interpretation is that if there were no representational relationship between the model and the real world it would be unclear how the model could teach us anything at all about a real-world target (Fumagalli 2016). In other words, if Schelling's model does not represent any feature of actual patterns of segregation, how could it produce a justifiable change in confidence regarding the necessary causes of real-world segregation? Presumably, the only way it could do so would be by latching onto some features of real-world patterns of segregation. As noted above, Schelling was trying, albeit informally, to connect the preference for not being in a minority assumed in the model with the real world. Hence, the problem might not be that there is no similarity between the model and the target, but that the model misrepresents its target in important respects, namely features of real-world cities and real-world migration processes.

This brings us to consider the second interpretation, according to which the model fails to represent its target. However, whether the respects in which the model is similar and in which it is not are relevant to explanation depends on what one takes the explanandum of Schelling's model to be. It could either be a specific occurrence,

namely segregation in a certain city, or the generic phenomenon of residential segregation, corresponding to target-directed modeling and generalized modeling, respectively. The mechanism Schelling's model aims to capture might have different degrees of importance for each explanandum. For example, if the explanatory claim concerns the actual pattern of segregation in a specific city, it is likely to be one among several difference-makers. The explanandum could also be the generic phenomenon of spatial segregation, and not just residential segregation. In this case, the mechanism Schelling identifies could even be the main explanation if it accounted for the probably few features that different kinds of spatial segregation (in cities, churches or restaurants, for example) have in common. Schelling himself points out that "[t]he analysis is so abstract that any twofold distinction could constitute an interpretation … the only requirements of the analysis is that the distinction be twofold, exhaustive, and recognizable" (Schelling 1978: 138). Each explanandum requires a different degree and a different kind of similarity. Hence, the model might fail to have the right similarity for providing a candidate explanation of one explanandum, but have the right similarity for another. Some of the arguments advanced to claim that Schelling's model does not provide a how-actually explanation point to the dissimilarity between the model and particular cases of residential segregation. Such arguments cast doubt on the model's capacity to explaining particular occurrences. The possibility that the model might succeed in picking out the actual causes or mechanisms of more general phenomena however remains unscathed.

On the third interpretation the reason why Schelling's model is not explanatory is that its representational adequacy has yet to be established. According to Reutlinger et al. (2017), for example, the model provides a how-possibly explanation because some of the assumptions, such as that agents know how many agents of each color live in their neighborhood, or that there are no social and economic factors, cannot be interpreted as being about explanatorily irrelevant factors *without further argument*. Such arguments would presumably concern the hypothesis that the unrealistic assumptions of the model are about explanatorily irrelevant factors, or that they do not matter to the result. This does not mean that

the model does not explain, however. It only implies that for the want of further arguments one is not entitled to believe in its explanation. This is precisely what the strategies of model verification I will discuss in Section 5 are supposed to do.

4.2 The Prisoner's Dilemma and the explanation of truces

My second example is Robert Axelrod's (1984) explanation of the 'live-and-let-live system' in the World War One trenches in accordance with the iterated prisoner's dilemma. What led Axelrod to hypothesize that the iterated prisoner's dilemma was applicable to this case was that truces occurred spontaneously and despite the pressures against them. In accordance with the PD, the structure of incentives makes a difference as to whether truces occur or not. Relying on Ashworth's historical analysis of WW1 warfare, Axelrod (1984: 75) identified features of the situation along the Western Front that made it an example of a PD. Northcott and Alexandrova (2015) argue that the PD is neither explanatory nor heuristically valuable when it comes to explaining the live-and-let-live system of WW1. They suggest that explaining a social phenomenon such as WW1 truces as such an instance amounts to claiming that "the *structure of the situation in conjunction with the actor's rationality* caused the outcome" (Northcott and Alexandrova 2015: 67; my emphasis). They acknowledge that there are similarities between the situation in WW1 and the one represented in the PD. However, they also claim that there are several differences between the two, which cast doubt on the PD's explanatory power. They further suggest that this is not the only reason to doubt that the PD explains the WW1 truces. In addition, the PD model does not address related explananda; the explanation comes after the fact; it is difficult to exclude the possibility that other games will suit the situation equally well; and it has not been demonstrated that the model's assumptions (perfect rationality and perfect knowledge) are satisfied. Given that Ashworth's original historical analysis identifies the actual causes of the truces, Northcott and Alexandrova (2015) conclude that it offers a better explanation than the one based on the PD.

I do not intend to challenge their claim that historical analysis offers a better explanation, or the more general one that the fruitfulness

of a certain way of understanding the social world should be evaluated in relation to other methods. What I wish to reconsider is their claim that the PD is not explanatory of WW1 truces. To this end, let us examine the first objection, namely that the differences between the situations represented in the PD and the WW1 truces cast doubt on the PD's explanatory power. On the basis of Ashworth's analysis, Northcott and Alexandrova conclude that features not present in the model were relevant in explaining the WW1 truces, hence the model does not include all the relevant difference-makers. This shows that the PD's explanation is partial, however, not that the PD does not explain. Moreover, as the PD is clearly not a model that directly represents WW1, it is plausible to interpret it as providing coarse-grained explanatory information. This may well be how Axelrod interprets it.

> The value of an analysis without [real-life complications] is that it can help clarify some of the subtle features…which might otherwise be lost in the maze of complexities of the highly particular circumstances in which choice must actually be made. It is the very complexity of reality which makes the analysis of an abstract interaction so helpful as an aid to understanding (Axelrod 1984).

The presence of differences alongside similarities is to be expected when a simple model is applied to a specific target. The similarities should be relevant and the differences irrelevant to the purpose at hand. For the model to provide coarse-grained explanatory information, it must be similar to the target only concerning the features that are not specific to the occurrence in question, but are shared by other systems. It might well be that Axelrod intended but failed to give a different, more complete explanation. Again, my claim here is not that the PD provides explanatorily relevant information, but rather that the presence of differences between the model and the target does not warrant the conclusion that the PD is not explanatory. Let us recall that this was not the only objection that Northcott and Alexandrova rose against the PD. So it might still be that even if the differences do not show that the PD is not explanatory, the other objections do. I will return to these in the next section since they pertain to the question of whether we are justified in believing that the PD does, in fact, provide explanatorily relevant information for WW1.

5 Increasing confidence in model-based explanations

Thus far, I have focused on the conceptual question of what it means for a model to be explanatory and argued that its answer depends partly on the attributes of the explanation for which it is being evaluated. But once we are clear on the kind of potential explanations that is our target, another question, the epistemological one, needs to be addressed, namely whether the model succeeds in picking out actual difference-makers. However, whereas a model either captures the right explanantia and hence it either explains or it does not, confidence in a model-based explanation is not an either-or matter, but a matter of degree.[5]

A variety of strategies of model verification contributes to building confidence in a model-based explanation. Some of these are empirical whereas others are not, even though non-empirical strategies are not sufficient in themselves to justify confidence in any given model-based explanation. In this section I show first that when properly understood derivational robustness analysis (a non-empirical strategy of verification) might be a better remedy to the problem of the sensitivity of the results to false assumptions than is sometimes thought. Next I argue that the difficulty of directly testing theoretical models coupled with casual empiricism need not lead to general skepticism about the explanatory power of economic models.

5.1 The role of robustness analysis

As pointed out in Section 2, one of the features of idealized economic models that is claimed to cast doubt on their explanatory power is that their results crucially depend on assumptions that are known to be false (e.g. Cartwright 2009). This is not the same as claiming that the model's explanantia do not satisfy the veridicality condition, however. Rather the problem is that modelers are not in a position to justifiably conclude that the model captures actual difference makers (i.e. the epistemological issue). Take, for example, Schelling's model. That the city has the shape of a checkerboard is obviously an

[5] When the evidence indicates that the explanantia are false, then we should conclude that the model does not explain. See however e.g. Bokulich 2009.

unrealistic assumption. The falsity of this assumption does neither imply the falsity of the model's explanantia nor that modelers are not justified in believing that the explanantia might be correct. For example, if the same result holds regardless of the assumed shape of the city, then the modeler's confidence in the explanatory power justifiably increases. The procedure by which economic modelers check which assumptions are crucial to their results is known as *robustness analysis*, the rationale of which is to show that a given falsity is not necessary to derive a result (Woodward 2006). This is not achieved by way of replacing the false assumption with a more realistic one, however, but by showing that the same result is obtained with a different yet still false assumption. Robustness analysis can show that the result depends on some false assumptions or that it does not. In either case, the modeler learns something useful, but only in the first case does confidence in the result legitimately increase.

Even though many philosophers concerned with economic modeling acknowledge the usefulness of robustness analysis, not all of them agree that it solves the problem of the falsity of assumptions. Several concerns have been raised: (i) many economists probe the robustness of the result against changes in only one or two assumptions; (ii) even if a robustness analysis is conducted, the results are seldom found to be robust to changes in assumptions; (iii) even when a robust theorem is found, its being robust does not guarantee its truth (e.g. Cartwright 2009, Odenbaugh and Alexandrova 2011, Reiss 2012, Woodward 2006). I do not intend to argue that all, most or many economic models are explanatory, and merely wish to show that the above claims (i-iii) stem in part from misinterpreting the scope of robustness analysis. My objective is to clarify the role that robustness analysis plays in increasing confidence in model-based explanations as well as in identifying which of a model's assumptions are crucial to which conclusions. These observations have already been made elsewhere (Kuorikoski, Lehtinen and Marchionni 2010, 2012). The reason to rehearse them here is that a clarification of the role of robustness analysis helps to sort out what the problems with economic modeling might be.

Kuorikoski, Lehtinen and Marchionni (2010, 2012) argue in their application of *derivational robustness analysis* to economic modeling that it provides a way of dealing with the unrealisticness of

tractability assumptions (see also Levins 1966, Wimsatt 1981, Weisberg 2006). Models that share a common structure but vary in their tractability assumptions are compared to identify a *robust theorem*, which states that the core assumptions shared by a group of models, call it C, lead to a certain result, P; that is, (C→P). The core assumptions C in potentially explanatory models are intended to represent the working of some difference-maker or mechanism and hence to be sufficiently realistic. Deriving the same result from models with the same core assumptions but different tractability assumptions increases confidence that the robust theorem is not an artifact of specific tractability assumptions. Derivational robustness analysis also helps to identify which modeling assumptions drive a given result. It is the latter assumptions that matter to the truth or falsity of the result.

Recall that the first concern about robustness analysis is whether economic modelers check the sensitivity of their results to enough changes in assumptions. Let us distinguish two interpretations of this claim: whether individual modelers probe the robustness of their results against only a few of their models' assumptions, or whether there is a large enough group of models that share the same set of substantive assumptions but have different tractability assumptions. Kuorikoski et al. (2010) interpret robustness analysis as a collective endeavor involving several modeling efforts. Hence, the fact that only one or a few assumptions are checked for robustness in a specific presentation of the model is not necessarily a good indication that robustness analysis is rarely conducted. Interpreted as a collective endeavor, it spans several modeling exercises. In Schelling's case, for example, efforts directed at exploring how results change with respect to different sets of assumptions are still on-going (cf. Ylikoski and Aydinonat 2014 and the references therein). Interpreted as a collective endeavor, robustness analysis turns out to be a rather widespread practice. Even so, whether it is carried out often enough to warrant general optimism in the explanation of economic models remains an open empirical issue. If it were not, the solution would be straightforward: robustness analysis should be conducted more often than it is now. A related concern is that some core assumptions of economic models are neither revised in the light of empirical evidence nor subjected to

robustness checks. Although this is a legitimate concern, many theoretical models work on the assumption of bounded rationality and/or imperfect information, which implies that at least some core assumptions are replaced with different ones. Moreover, even if it were the case that some core assumptions are never replaced, it would not rule out the possibility that economic models explain, nor would it call into question the epistemic power of robustness analysis. Again, the solution would be more rather than less robustness analysis.

The second concern is that in economics cases abound in which robustness analysis is carried out but does not yield a robust theorem. As explained above, if a result is not robust to changes in tractability assumptions, then confidence in the model should not increase. The mechanism represented in it may be a difference-maker, but there is no justification for believing that it is (Hausman 2013: 253). If robustness analysis in economics turned out to fail on a global scale, it would indeed cast doubt on the explanatory potential of economic models in general. Establishing the frequency of robust theorems in economics lies beyond the scope of this paper. Nevertheless, at least a few theorems seem to be robust to changes in tractability assumptions, Schelling's result being a case in point (Ylikoski and Aydinonat 2014: 23). Moreover, as Hands (2016) argues, the fate of models is at least sometimes determined by their performance in robustness checks.

The last objection to robustness analysis, that it cannot alone deliver the truth, derives from overstating its epistemic import. In the absence of empirical support for any of the assumptions, Cartwright is right: the inductive leap from the model to the world is a very bad bet. This is precisely why some assumptions, at least the core assumptions that are shared across models, should have some degree of empirical support. Robustness analysis does not guarantee the truth, but it is nevertheless epistemically valuable: by means of identifying which assumptions are required to derive the results it indicates which parts of the model must be similar enough to the target to allow the model to explain (some feature of) it.

5.2 Empirical strategies of model verification

Robustness analysis facilitates the identification of assumptions that are crucial to the model's result. It does not justify the belief that the model provides explanatorily relevant information in the absence of empirical support for any of its assumptions. If we knew all the crucial assumptions were true, then we could confidently conclude that the explanation based on those assumptions is the actual one. Yet, only in a small minority of cases can it be confidently concluded on the basis of the evidence that a given model-based explanation is an actual explanation, or ruled out that it is not. The situation that is most commonly faced is one of uncertainty regarding whether a given model picks out actual difference-makers (see also Hausman 2013). It is therefore useful to think of model-based explanations as lying on a continuum ranging from potential explanations for which justification is weak (provided they have a certain degree of plausibility in the light of background knowledge) to potential explanations for which justification is stronger.

Increasing confidence in a model-based explanation requires support for the hypothesis that all the features that should be similar between a model and a target given a specific explanatory inference are, in fact, similar, and that the differences are irrelevant.[6] As others (e.g. Sugden 2000, Weisberg 2013) have already noted, learning with models involves analogical reasoning based on the following structure:

(1) The model, *m*, and the target *t*, are relevantly similar

[6] The claim that models represent in virtue of similarity relations to their world targets is far from uncontroversial, especially because of the difficulties in determining whether relevant similarity is present and whether it is sufficient for making inferences from the model to the target (see e.g. Weisberg 2013 for an attempt to give formal treatment to similarity judgments, and Parker 2015 for objections). For the purposes of this paper, I simply assume that such difficulties can be resolved, although I do not attempt to do so. This assumption is legitimate given that my aim is to find out whether there is anything specific to economic models that sets them apart from models in other fields, and not to offer a defence of the representationalist account of scientific models. In addition, many of the sceptical arguments I examine suffer from (a lack of) similarity between models and targets.

(2) ϕ is true of m

(3) Therefore, ϕ is true of t.

Premise (1), the base of the analogy, is the hypothesis on which evidence is brought to bear. What can be inferred and how secure the inference is depend on the kind and amount of evidence for (1), whereas relevant similarity is dependent on the kind of explanation of interest. If, say, preferences for not being in a minority make a difference to whether segregation occurs in the model or not, then whether this is an actual difference-maker of real-world segregation depends on whether the model and the target are relevantly similar. Given the uncertainty about the base of the analogy, there is also uncertainty about whether conclusion (3) is the case (cf. Norton 2011).

Whereas non-empirical strategies such as robustness analysis help to refine judgments regarding premise (2), that is, regarding what are the crucial assumptions for a given result, there are different ways of increasing confidence in relevant similarity between model and target (premise 1). One strategy consists in testing the model's results against data, either observationally or experimentally. Both the (iterated) PD and Schelling's models have been subjected to several empirical and laboratory tests, with mixed results.[7] As noted above, however, one of the peculiarities of economics is that empirical tests are rarely direct tests of the theoretical models, those being typically done via only loosely connected empirical models. Model testing is not the only way of bringing empirical evidence to models, however. Here I will comment only briefly on two other strategies, namely independent support for assumptions and mechanistic tracing, which complement standard model-testing techniques (Rodrik 2015). Combinations of strategies provide stronger support for the model than each strategy in isolation (e.g. Lloyd 2010, Lehtinen 2016).

Independent support for the model's assumptions brings evidence to bear on the hypothesis that there is a relevant similarity between the assumptions and features of the (specific or generalized) target.

[7] The experimental literature on the PD is massive. For overviews of experiments on specific versions see, for example, Ledyard 1995 and Holt et al. 2015. For tests of the Schelling-style segregation model see, for example, Clark 1991 and Tsvetkova et al. 2016.

For example, if the PD is to contribute to explaining the occurrence of the WW1 truces, then what must be checked is whether the conditions concerning incentives, information and repetition the PD identifies as making truces possible were in fact present during WW1. Axelrod's use of Ashworth's historical data aims precisely to establish whether the PD model's key assumptions have similar features to the WW1 live-and-let-live system relative to features they do not share. The distinction between fine-grained and coarse-grained explanatory information indicates what features should be shared: in the case of coarse-grained explanatory information things such as the timing of the truces, which is likely to differ across situations, are irrelevant. By way of contrast, in the case of Schelling's model it has been found that in actual cases of racial segregation in the US people vary in their racial preferences, and that many have discriminatory preferences. The implication is that the contribution of Schelling's mechanism to producing real patterns of segregation in US cities is probably small, but it does not demonstrate that it is not a difference-maker and hence that the model does not explain. As should be clear from Section 3, we need not require that for a model to be explanatory it must pick out the only or the main difference-makers.

Mechanistic tracing involves marshaling evidence for hypotheses about the similarity of mechanisms. As a clear and intuitive illustration, I borrow one of Dani Rodrik's examples (2015: 100). The *Dutch disease model* accounts for why the discovery of a natural resource can harm a country's economic performance. The postulated mechanism is as follows: the country's exchange rate appreciates as a result of its resource abundance, making the manufacturing sector less profitable, which in turn has negative repercussions on the whole economy. If a country that has experienced a resource boom is now in an economic downturn, and there is evidence that the manufacturing sector became less profitable during the intervening stages, there is reason for increased confidence in the model's mechanistic claim. In the case of Schelling's model it is possible to find evidence that in, say, a specific neighborhood in which the majority has a preference for not being in a minority, people typically move to another neighborhood only when the number of neighbors of the other group reaches a certain threshold. Ceteris paribus, mechanistic tracing

provides more secure evidence the more mechanistic details the model includes that can be compared with the target. In Schelling's model the mechanism is described sparsely and abstractly, and hence its similarity to real-world mechanisms would only warrant equally sparse and abstract explanations (cf. Steel 2013).

6 From potential to actual model-based explanations

In Section 4 I have discussed the explanatory power of Schelling's model and of the PD while bracketing the epistemological issue of whether we are justifying in believing that they actually are explanatory. This is what we examine here starting from Schelling's model.

We have seen that what counts as relevant and irrelevant similarity depends on the explanatory task for which we are assessing the model. Some of the arguments advanced by those who doubt the explanatory power of Schelling's model, such as the model does not resemble any real city, demonstrate that the base of the analogy is to be rejected when it comes to explaining particular cases of residential segregation. When it comes to explaining the generic phenomenon of spatial segregation, however, the possibility remains open that the base of the analogy is only in need of further justification. In other words, the problem in this case would be that it has not yet been established whether the crucial assumptions in Schelling's model sufficiently resemble the actual phenomenon of spatial segregation for us to legitimately believe that Schelling's mechanism is among its actual difference makers.

Clearly, one could demand that only models with established similarity should be deemed explanatory. I have no issue with taking models for which such a warrant is missing as potentially explanatory rather than genuinely explanatory. What I object to is the conclusion that they are *not* explanatory, and hence that in order to make sense of why such models are valued, one should look for some other contribution they make that is not explanation. If one thinks of model-based explanations as lying on a continuum from potential to actual, the characteristic of Schelling's original model is that it provides a partial explanation upon which only casual empirical considerations concerning the behavioral assumptions are brought to bear. It may be that not much more than this is needed given the sparseness of the model in terms of

mechanistic details—even though its explanatory power would then turn out to be limited along this dimension (cf. Bokulich 2014).[8]

My diagnosis of the problems with the PD explanation of the WW1 truces proceeds along similar lines. I argued above that what counts as explanatorily relevant and irrelevant differences depends on the explanatory information the PD is taken to provide about the WW1 truces. Yet, this does not suffice to legitimate the belief that the PD provides relevant explanatory information about the WW1 truces. In fact, the observation of differences between the PD and the WW1 truces was not the only reason why Alexandrova and Northcott questioned the model's explanatory power. They also claim that the PD does not address related explananda; that the explanation is after the fact; that it is difficult to exclude the possibility of other games fitting the situation equally well; and finally, that it is not demonstrated that the model's assumptions (perfect rationality and perfect knowledge) are satisfied. My contention is that these objections justify not having great confidence in the PD explanation of the WW1 truces, but do not demonstrate that the PD does not explain. First, suppose that the PD model accounted for related explananda: why should this be a reason to believe that it explains the original explanandum, namely why the truces occurred in the first place? Its capacity to explain other features of the WW1 truces could indicate that the PD explanation has a broad scope or, depending on one's theory of confirmation, it might count as stronger evidence in favor of the model. It does not necessarily have implications as to whether the model explains the original explanandum, however. A similar point applies to the two subsequent objections. If the model provided novel predictions, or was the only game that fit the situation, then support for the PD explanation would be stronger (to different degrees depending on one's preferred theory of confirmation), but their absence does not imply that the PD does not explain. Finally, what about the fact that the model's assumptions of perfect rationality and perfect knowledge are not satisfied in the WW1 truces? If it were indeed the case that such assumptions were

[8] This is compatible with the model giving a how-possibly explanation in the sense proposed by Grüne-Yanoff (2013), in addition to, or instead of, the model actually explaining an aspect of the generic pattern of segregation.

not even approximately true of the WW1 situation, and if truces did not emerge spontaneously if individuals were less than rational and fully informed, then this is the one case in which we should not have much confidence that the PD captures the actual difference-makers of the spontaneous emergence of the truces.

Before giving my conclusions I should address a general objection against the claim that models such as Schelling's and the PD can be explanatory. It could be argued that if the conditions under which the causes identified in the model make a difference are identified through detailed historical analysis, for example, then it is not the model that does the explaining, but the detailed historical analysis. This is the position Alexandrova and Northcott (2013) seem to favor in regarding models as *open formulas*, in other words as simple heuristics for the construction of templates to be filled in by means of empirical investigation (see also Alexandrova 2008). The issue under contention here pertains to the model's failure to provide a full specification of the causal hypothesis. I agree that most theoretical models fall short of providing such a full specification, but this does not mean that they are merely heuristic devices or sources of inspiration. It is one thing to empirically warrant the causal hypothesis independently of the route through which it was formulated, and another to warrant it by way of verifying the similarity between the model and the target. In the latter case the model plays an indispensable role not only in the process of formulating but also in that of justifying the causal hypothesis. Northcott and Alexandrova are right in claiming that the application of the PD (and of Schelling's model for that matter) to specific instances shows that it is hardly sufficient to indicate how various contingent factors interact to bring about a given outcome. Such information must come from elsewhere. This does not necessarily make the model non-explanatory, however. It indicates that generic and abstract explanatory claims are of limited help in specific applications, but this holds regardless of whether the claims are based on models, experiments or any other methodology.

7 Concluding remarks

I have argued that in order to assess the explanatory power of idealized models it is useful to keep two sets of issues clearly separate: what conditions should a model satisfy in order to count as explanatory and whether the model actually satisfies those conditions. Whereas the requirement that the explanantia be true sets explanatory models from non explanatory models apart, the typical situation economic modelers confront is one of epistemic uncertainty about whether the requirement is satisfied. A combination of empirical and non-empirical strategies of verification can be deployed to increase confidence in a given model-based explanation, but given some of the peculiarities of economics such strategies seldom provide conclusive evidence that a model-based explanation is adequate. Making clear that model-based explanations in economics lie on a continuum between being potential and probably, or very probably actual helps in diagnosing more accurately what, if anything, is wrong with theoretical modeling in economics. Possibilities include the following: whether some of the empirically questionable assumptions of economic models are never replaced by more realistic assumptions nor subjected to robustness analysis (e.g. Cartwright 2009, Odenbaugh and Alexandrova 2011); whether few robust theorems are discovered, or at least too few to justify the extent of the current devotion to building models that differ only on a few assumptions (e.g. Reiss 2012); and finally, whether most economists are merely content with very weak empirical support for very general claims (e.g. Northcott and Alexandrova 2015). What I hope to have shown is that the above are open empirical questions concerning the practice of economic modeling, which might be obfuscated by framing the issue in terms of truth and falsity of the explanantia as it is sometimes done. Establishing by empirical means which of these problems beset economic modeling would make it easier to find ways of remedying them.[9]

[9] This research was financially supported by the Academy of Finland. I would like to thank Emrah Aydinonat, Till Grüne-Yanoff, Francesco Guala, Maria Jimenez Buedo, Luis Mireles Flores, Federica Russo, and two anonymous referees for their valuable comments on previous versions of the paper. Some of the ideas expressed here have been presented at the conference on Causality and Modeling in the Sciences (Madrid), the workshop on Explanation, Normativity

Caterina Marchionni
Practical Philosophy & TINT
Faculty of Social Sciences
University of Helsinki
P.O. Box 24, 00014 Helsinki
Finland

References

Alexandrova, Anna. 2008. Making models count. *Philosophy of Science* 75: 383–404.
Alexandrova, Anna; and Northcott, Robert. 2013. It's just a feeling: why economic models do not explain. *Journal of Economic Methodology* 20(3): 262–8.
Axelrod, Robert. 1984. *The Evolution of Cooperation*. Penguin.
Aydinonat, Emrah. 2008. *The Invisible Hand in Economics: How Economists Explain Unintended Consequences*. London/New York: Routledge.
Backhouse, Roger. 2007. Representation in economics. In *Measurement in Economics: A Handbook*. Ed. by M. Boumans. Elsevier: 135–52.
Basso, Alessandra; Chiara Lisciandra; and Caterina Marchionni. 2017. Hypothetical models in social science. In *Springer Handbook of Model-Based Science*, ed. by Magnani and Bertolotti. Springer: 413–33.
Bokulich, Alisa. 2014. How the tiger bush got its stripes: 'how possibly' vs. 'how actually' model explanations. *The Monist* 97(3): 321–38.
Bokulich, Alisa. 2009. Explanatory fictions. In *Fictions in Science: Philosophical Essays on Modeling and Idealization*, ed. by M. Suárez. London: Routledge, 91–109.
Clark, W.A.V. 1991. Residential preferences and neighbourhood racial segregation: a test of the Schelling segregation model. *Demography* 28 (1): 1–19.
Cartwright, Nancy. 1999. The vanity of rigour in economics: theoretical models and Galileian experiments. In *The 'Experiment' in the History of Economics*,ed. by P. Fontaine and R. Leonard. Routledge, 135–53.
Cartwright, Nancy. 2002. The limits of casual order, from economics to physics. In *Fact and Fiction in Economics*, ed. by U. Mäki. Cambridge: Cambridge University Press, 137–51.
Cartwright, Nancy. 2009. If no capacities then no credible worlds. But can models reveal capacities? *Erkenntnis* 70: 45–58.
Fumagalli, Roberto. 2016. Why we cannot learn from minimal models. *Erkenntnis* 81 (3): 433–55.
Garfinkel, Alan. 1981. *Forms of Explanation*. New Haven: Yale University Press.
Grüne-Yanoff, Till. 2009. Learning from minimal economic models. *Erkenntnis* 70: 81–99.

---

and Uncertainty in Economic Modelling (LSE), and the workshop on Models and Explanation in Economics (Innsbruck). I would like to thank the participants at these events, and especially Robert Sugden, for insightful discussions.

Grüne-Yanoff, Till. 2013. Appraising models nonrepresentationally. *Philosophy of Science* 80: 1–12.

Hands, Wade. 2016. Derivational robustness analysis, credible substitute systems and mathematical economic models: the case of stability analysis in Walrasian general equilibrium theory. *European Journal for the Philosophy of Science* 6(1): 31–53.

Hausman, Dan. 2013. Paradox postponed. *Journal of Economic Methodology* 20(3): 250–4.

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

Hindriks, Frank. 2006. Tractability assumptions and the Musgrave-Mäki typology. *Journal of Economic Methodology* 13: 401–23.

Holt, Charles; Johnson, Cathleen; and Schmidtz, David. 2015. Prisoner's Dilemma experiments. In *The Prisoner's Dilemma*,ed. by M. Peterson. Cambridge University Press: 243–64.

Kirkham, Richard L. 1992. *Theories of Truth. A Critical Introduction*. The MIT Press.

Kuorikoski, Jaakko; Lehtinen, Aki; and Marchionni, Caterina. 2010. Economic modelling as robustness analysis. *British Journal for the Philosophy of Science* 61: 541–67.

Kuorikoski, Jaakko; Lehtinen, Aki; and Marchionni, Caterina. 2012. Robustness analysis disclaimer: please read the manual before use. *Biology and Philosophy* 27 (6): 891–902.

Jackson, Frank; and Pettit, Philip. 1990. Program explanation: a general perspective. *Analysis* 50: 107–17.

Lehtinen, Aki. 2016. Allocating confirmation with derivational robustness. *Philosophical Studies* 173 (9): 2487–509.

Ledyard, John. 1995. Public goods: a survey of experimental research. In *The Handbook of Experimental Economics*, ed. by J. Kagel and A. Roth.Princeton, NJ: Princeton University Press, 111–94.

Levins, Richard. 1966. The strategy of model building in population biology. In *Conceptual Issues in Evolutionary Biology* (1st ed.), ed. by E. Sober. Cambridge, MA: MIT Press: 18–27.

Lloyd, E. 2010. Confirmation and robustness of climate models. *Philosophy of Science* 77(5): 971–84.

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

Mäki, Uskali. 2013. On a paradox of truth, or how not to obscure the issue of whether explanatory models can be true. *Journal of Economic Methodology* 20(3): 268–79.

Marchionni, Caterina. 2008. Explanatory pluralism and complementarity: from autonomy to integration. *Philosophy of the Social Sciences* 38: 314–33.

Marchionni, Caterina. 2013. Playing with networks: How economists explain. *European Journal for Philosophy of Science* 3(3): 331–52.

Northcott, Robert. 2013. Degree of explanation. *Synthese* 190 (15): 3087–105.

Northcott, Robert; and Alexandrova, Anna. 2015. Prisoner's Dilemma doesn't explain much. In *The Prisoner's Dilemma*, ed. by M. Peterson. Cambridge University Press, 64–84.

Norton, J. 2011. Analogy. Unpublished draft, University of Pittsburg. http://

www.pitt.edu/~jdnorton/papers/Analogy.pdf.
Odenbaugh, Jay; and Alexandrova, Anna. 2011. Buyer beware: robustness analyses in economics and biology. *Biology and Philosophy* 26: 757–71.
Reiss, Julian. 2012. The explanation paradox. *Journal of Economic Methodology* 19 (1): 43–62.
Rohwer Yasha. and Rice Collin. 2016. How are models and explanations related? *Erkenntnis* 81 (5): 1127–48.
Rice, Collin. 2015. Moving beyond causes: optimality models and scientific explanation. *Noûs* 49(3): 589–615.
Rodrik, Dani. 2015. *Economics Rules. Why Economics Works, When It Fails, and How To Tell The Difference*. Oxford University Press.
Reutlinger, Alexander; Hangleiter, Dominik; and Hartmann, Stephan. 2017. Understanding with (toy) models. *British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axx005
Schelling, Thomas. 1978. *Micromotives and Macrobehavior*. London: W.W. Norton.
Steel, Daniel. 2013. Mechanisms and extrapolation in the abortion-crime controversy. In *Mechanisms and Causality in Biology and Economics*, ed. by H-K Chao, S-T Chen and R. Millstein. Springer, New York: 185–206.
Sugden, Robert. 2000. Credible worlds: the status of theoretical models in economics. *Journal of Economic Methodology* 7: 1–31.
Tsvetkova, Milena; Nilsson, Olof; Öhman, Camilla; Sumpter, Lovisa; and Sumpter, David. 2016. An experimental study of segregation mechanisms. *EPJ Data Science* 5: 4 DOI: 10.1140/epjds/s13688-016-0065-5.
Weisberg, Michael. 2006. Forty years of "The Strategy": Levins on model building and idealization. *Biology and Philosophy* 21(5): 623–45.
Weisberg, Michael. 2013. *Simulation and Similarity*. Oxford University Press.
Wimsatt, William. 1981. Robustness, reliability, and overdetermination. In *Scientific Inquiry and the Social Sciences*, ed. by M. Brewer and B. Collins. San Francisco: Jossey-Bass: 124–63.
Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
Woodward, James. 2006. Some varieties of robustness. *Journal of Economic Methodology* 13: 219–40.
Ylikoski, Petri; and Aydinonat, Emrah. 2014. Understanding with theoretical models. *Journal of Economic Methodology* 21(1): 19–36.
Ylikoski, Petri; and Kuorikoski, Jaakko. 2010. Dissecting explanatory power. *Philosophical Studies* 148 (2): 201–219.