Erkenn (2009) 70:81–99
DOI 10.1007/s10670-008-9138-6

ORIGINAL PAPER

Learning from Minimal Economic Models

Till Grüne-Yanoff

Received: 15 April 2008 / Accepted: 1 October 2008 / Published online: 24 December 2008

Abstract It is argued that one can learn from minimal economic models. Minimal models are models that are not similar to the real world, do not resemble some of its features, and do not adhere to accepted regularities. One learns from a model if constructing and analysing the model affects one's confidence in hypotheses about the world. Economic models, I argue, are often assessed for their credibility. If a model is judged credible, it is considered to be a relevant possibility. Considering such relevant possibilities may affect one's confidence in necessity or impossibility hypotheses. Thus, one can learn from minimal economic models.

1 Introduction

Do we learn from economic models, and if so how? In order to answer this question in the positive most philosophical accounts impose strict conditions. In particular, they require that models be linked to the real world, for example through similarity or partial resemblance relations, or through their adherence to natural laws. Yet it seems that economic modellers often do not heed these conditions when constructing and evaluating their models. It is therefore worth investigating models that are *minimal* in the sense that they do not satisfy these world-linking conditions, and finding out whether one may learn from them.

I argue in this paper that one may indeed learn from minimal models because they might affect one's confidence in impossibility hypotheses about the world. The paper starts with an account of such models, and proceeds to clarify what learning from a model means. The discussion then turns to Sugden's (2000; this issue) account of modelling, which is close in many ways to the one presented here. I will

T. Grüne-Yanoff (✉)
Helsinki Collegium of Advanced Studies, University of Helsinki, Fabianinkatu 24, P.O. Box 4, 00014 Helsinki, Finland
e-mail: till.grune@helsinki.fi

show that the conclusions he envisages are too strong, and that therefore his argument does not account for how one justifiably learns from minimal models. Taking his suggestion as my starting point, I rather analyse the notion of credibility and show which inferences a credible model licenses. On the basis of the results I then argue that we learn from minimal models because they may affect our beliefs about what is impossible or necessary in the real world.

2 Minimal Models

Theoretical economists are modellers. They hope that constructing, manipulating and analysing models will reveal something of interest about the world. Philosophers of science have described such modelling activity as the construction of 'representations-as' (Hughes 1997, p. 331), as a medium for 'surrogate reasoning and inference' (Suárez 2004, p. 767), and as a 'strategy of indirect representation' (Godfrey-Smith 2006, p. 730). Although differing in detail, there is a consensus in all these views that modelling consists in the construction of artificial systems that act as stand-ins for real-world objects or systems, and that are analysed in their stead.

There is considerable disagreement, however, concerning the conditions that models must satisfy in order to function as stand-ins. The notions of surrogacy and representation suggest that a model stands in relation to a particular real-world target, that one learns about this target by examining the model, and that the quality of this model-target relationship determines what can be learned (cf., for example, Mäki, this issue, section 2.1). Yet for many models no specific target is identified (Cartwright, this issue, section 1; Knuuttila, this issue, section 6). Nor, as I will argue in this paper, do models commonly adhere to natural laws. Many economic models thus lack the world-linking relations that philosophers of science impose on them.

The strategy followed in this paper, therefore, is to eschew these conditions, and instead to characterise economic models by the salient features of their common scientific usage. The resulting *minimal model* is then examined in the light of the potential learning effects it may offer.[1]

The two salient features of models used in economics are their surrogate nature and their internal dynamics. They are used as *epistemic surrogates* in the sense that modellers focus on them in place of something that is the eventual object of their interest. The modellers construct, manipulate and analyse models, but their ultimate rationale for doing so is to learn not about the model, but about something else. This eventual object of interest need not be a particular. It may be an abstract feature of a class of phenomena, or even a vague idea about such a feature, without any

[1] Suarez defends a 'deflationary or minimalist attitude and strategy towards the concept of scientific representation' that is related to the minimal model presented here (Suárez 2004, p. 770). However, he suggests that representational force and allowing inferences are necessary conditions for scientific representations. It seems to me that both conditions are rather matters of degree, and thus only function as comparative but not exclusionary criteria. The minimal-model account presented here does not lay claim to any necessary criteria.

reference to the real-world phenomena that exhibit it. For example, economists often investigate equilibria in their models. It is clear from their writings that they want to learn about equilibrium states outside of the specific models they are using, yet they hardly ever identify the class of concrete situations that they think may exhibit such a feature. However, a model inadvertently exhibits a concrete if imaginary situation, typically a system of commodities, prices, consumers and sellers. A model's surrogate function is thus much wider than merely *replacing* a concrete real-world system with a surrogate system. Rather, the relation between the model and the real world may be left undetermined, and still the model is used to facilitate learning about the world.

In a way, all theories are surrogates: their immediate focus is on the representation and not the represented, even if the latter is the goal of theorising. Yet some theorists aspire to foster close relations between their representation and the represented. For example, they develop theoretical representations of a real system from in-depth data analysis, arguing for each abstraction or simplification that distinguishes the representation from the data. They aspire to 'direct representations' (Weisberg 2007) of real-world systems derived from observations about concrete phenomena, which they closely resemble albeit in an abstract and simplified form. The practice of modelling differs from those modes of theorising in the *absence* of such a purported or actual derivation from real-world systems. Instead, modelling starts with an asserted surrogate relation, thus being less demanding than the other modes (Godfrey-Smith 2006; Weisberg 2007).

In the minimalist spirit of this paper, I propose that the difference between such 'directly representing' theories and models may go beyond a difference in their construction practice. Minimal models are assumed to lack any similarity, isomorphism or resemblance relation to the world, to be unconstrained by natural laws or structural identity, and not to isolate any real factors. I then ask whether minimal models may still function as epistemic surrogates—i.e. whether one could still learn anything about the world from them.

The second salient feature of model use in economics is the model's dynamic aspect, which allows the modeller to manipulate it in some way and subsequently to obtain some 'results'. What this means in detail depends on the nature of the surrogate system. Users of material surrogate systems such as laboratory rats, wind tunnels, or hydraulic systems are able to exploit the causal mechanisms that operate within the system's boundaries. Their intervention leads to a change in the material system, which constitutes a model 'result'.

Current economic modelling almost exclusively uses formal systems, which in contrast to material systems have their dynamic aspect specified through explicit assumptions. These dynamic aspects may be very simple: the result may be derived from resetting a parameter value, or from 'shifting the curve' of a graph, for example. In other cases the derivation may be fully formalised in a formal system: a system of propositional calculus, for example, determines its deductive conclusions through specified rules of inference; an economic supply-demand system is resolved by means of linear programming; and an agent-based computer simulation is computed on the basis of the agents' specified 'behavioural rules'.

A formal structure is an important constituent of economic models, but not the only one. Take, for example, game-theoretic models: the same formal structure (usually a set theoretical object) of a game yields different results depending on how it is interpreted. If the players are human beings the result will be different from cases in which they are the 'selves' of the same person, or bees. If payoffs are utility numbers, the model result again will be different from interpretations according to which they are sums of money or measures of evolutionary fitness. The results that a model yields thus depend not only on its formal characterisation, but also on the economic interpretation of this formal structure (Rubinstein 2001; Grüne-Yanoff and Schweinzer 2008). Economic models consist of both a formal structure and its interpretation (cf. also Gibbard and Varian 1978).

Nevertheless, a model's interpretation must be distinguished from the systems that modellers hope eventually to learn about. A close look at economic models reveals that utility functions are routinely called 'consumers', 'bidders' or 'buyers', probability functions 'information', and linear programs 'markets'. The way theoretical modellers name components of the formal structure shows that they think of the model as a concrete situation—yet not a situation of the real world. They interpret formal structures not as descriptions of the real world but as describing 'parallel worlds' (Sugden 2000, p. 25), which exhibit familiar features of the real world but may not be identifiable with any of its particulars. Thus, economic models consist of both a formal structure and an interpretation of this structure as an imaginary scenario or world—and both of these components together serve as epistemic surrogates from which modellers hope to learn about their ultimate target.

In all economic models one finds the combination of a constructed formal structure and its interpretation as an imaginary, concrete economic system that is ready to be manipulated in its variables and parameters, and ready for its results to be analysed. I call this characterisation the *minimal model* because its representational function as a surrogate is merely declared, and no further claims are made about the truth of its assumptions, the epistemic status of the principles used in its construction, or the similarity of its economic interpretation (or parts of it) with the real world. This paper investigates whether one can learn from minimal models about the world.

3 Learning

Economics, in the self-conception of most economists as well as in its public image, is a science. Economists purport to predict, to explain, and to give policy advice. Public and private decision makers draw on the results, insisting on the scientific status of economics. As a science it is committed to certain epistemological goals. The practices it employs must therefore be justified in the light of these goals. In particular, in order to justify modelling practice one must provide epistemological rather than pragmatic or sociological reasons. I propose that the epistemological goals of modelling are directed towards *learning*.

Learning from models is an effect of their use on our knowledge. Models affect different domains of knowledge—about the model itself, about the theories from

which it is constructed, or about the world. It is the last domain I will focus on here because this kind of knowledge is necessary in order to explain, predict, or give policy advice. Learning, I therefore suggest, is constituted by a *justified change in confidence in certain hypotheses about the world.*

The use of models for learning purposes differs from other types of usage. In particular, it contrasts with their use in the development of new hypotheses. For example, game theory facilitates the construction of various types of game models, such as the Chicken game or the Prisoners' Dilemma. These models allow the expression of a hypothesis about the world—such as 'this situation is a Prisoners' Dilemma'—with a degree of precision that is difficult to reach without the help of such a model. Game models thus provide a language in which hypotheses can be formulated more precisely (Ginits 2000, p. xxviii). Beyond increasing the precision of hypotheses, models also help in structuring claims about the world. For example, game theory offers taxonomies of game models ranging from simple dual distinctions—such as cooperative versus non-cooperative—to a comprehensive categorisation of game types according to the inequalities between their payoffs. In considering the different types of games researchers are able to explore various hypotheses, and to make a more structured choice.

Such heuristic uses of models are important, but they do not constitute learning: if model use merely contributes to the *formulation* of the hypothesis, then no learning takes place. Learning requires that the model effects justified changes of our *confidence* in the hypothesis, or if the model itself gave rise to its formulation that it forces us to form a belief about this newly-formulated hypothesis based on consideration of the model itself.

Learning from models became prominent as a topic through Morgan and Morrison's (1999) work. However, closer analysis of their texts reveals that their notion of learning from model construction and manipulation does not fit the more stringent characterisation proposed here. According to Morgan and Morrison, models primarily teach one about the model world, and about theories. When it comes to learning about the world, Morgan reverts to the more standard notion that models must resemble the real world: 'if we want to use models to learn about the world, the model needs to map onto the real world' (Morgan 1999, p. 366). The minimalist account proposed in this paper argues against the necessity for resemblance (or other world-linking properties) in learning from models.

4 Stronger Model Requirements

Can we learn anything about the real world from minimal models? The standard answer from philosophers of science and methodologists alike is no: a model must satisfy properties beyond those of the minimal model in order to be useful in such learning. In particular, these additional properties are supposed to secure, in one way or another, a connection between the model and the world. Prominent examples of approaches that stipulate such properties include Cartwright's capacity account, Giere's (1988) similarity account, and Mäki's isolation account.

Cartwright argues that we learn from models if they function as Galilean thought experiments. Their aim is to replicate real Galilean experiments in which real-world systems are designed in such a way that the cause in question operates 'on its own' or 'without impediment'. In the laboratory this cause produces its effects by 'exercising the capacity as dictated by Nature's principles', whereas in a model, 'we produce the effects by deduction from the principles we adopt in the model' (Cartwright, this issue, section 2). Thus, Cartwright agrees with the characterisation of models as containing a formal system that facilitates the derivation of the results. Crucially, however, she adds the requirement that this formal system be based on confirmed capacity principles. Whether a modeller learns from a model thus largely depends on the principles she has chosen, especially their truth, stability and separability (Cartwright 1999).

According to Giere, the relations between scientific models and systems in the world are mediated by *theoretical hypotheses* (cf. Giere 1988, p. 80). Theoretical hypotheses make the claim that a designated real system is similar in specified respects and to specified degrees to a proposed model. Applied to economics, typical examples of such hypotheses assert that 'some actual economic objects, at least to some degree of approximation, constitute economic equilibrium systems'.[2] According to Giere's specified similarity account, we learn from models if the theoretical hypothesis that mediates between the model and the system in question is true—i.e. if the model is actually similar to the system in the respects and to the degree specified in the theoretical hypothesis.[3]

This also holds for Mäki's account. He sees models as partial representations: instead of representing a whole system, a model only represents parts of a real system in an isolating model environment. In order for us to learn from models about the world, this partial representation relation must turn out to be a partial resemblance—i.e. the model components must resemble the relevant parts of the real system. Thus, 'in the real world, the assumed isolation does not exist, whereas the isolated force does exist' (Mäki 2004, p. 1727), and the successful model is 'true, i.e. nothing-but-true, about those parts' (Mäki 1994, p. 159). While Mäki qualifies the resemblance with various pragmatic considerations (Mäki, this issue, section 2.2), it remains the case that similarity between the model and the world is a precondition for learning from the model—modelling attempts end in 'weak failure' if 'the model does not resemble the target in appropriate ways' (Mäki, this issue, section 4).

Any of these additional criteria may be sufficient: if a model satisfies one of them, one arguably learns from it about the world. Yet if we look more closely at theoretical modelling practice we will observe that economic modellers generally do not argue for a link between the model and the world when constructing the model, nor do they check for it once the model has been constructed.

In particular, when economic theorists describe the process of constructing a theoretical model they stress the role of creativity, playfulness and imagination—

[2] Cf. Hausman (1992, p. 75).

[3] Giere has since proposed an alternative, pragmatic account that does not rely to this extent on similarity (Giere 2004). His previous account, however, still attracts attention, particularly among scientists. It therefore seems worthwhile discussing it here.

but they do not mention the importance of well-confirmed theoretical principles, of data analysis, or of any other well-argued link to actual situations. Schelling gives a good illustration of this construction process when he describes how he came to develop his famous checkerboard model:

> Sometime in the 1960s, I wanted to teach my classes how people's interactions could lead to results that were neither intended nor expected. I had in mind associations or spatial patterns reflecting preferences about whom to associate with in neighbourhoods, clubs, classes, or ballparks, at dining tables … I found nothing I could use [as illustrative material], and decided I'd have to work something out for myself. One afternoon, settling into an airplane seat, I had nothing to read. To amuse myself I experimented with pencil and paper. I made a line of pluses and zeros that I had somehow randomized, and postulated that every plus wanted at least half its neighbors to be pluses and similarly with zeros. Those that weren't satisfied would move to where they were satisfied … At home I took advantage of my son's coin collection … I spread [the coins] out in a line, either in random order or any haphazard way, gave the coppers and the zincs their own preferences about neighbors, and moved the discontents – starting at the left and moving steadily to the right – to where they might inject themselves between two others in the line and be content. The results astonished me … I experimented with different sizes of "neighbourhoods" … A one-dimensional line couldn't get me very far. But in two dimensions it wasn't clear how to intrude a copper or a zinc into the midst of coppers and zincs. I mentioned this problem to Herb Scarf, who suggested I put my pennies on a checkerboard, leaving enough blank spaces to make search and satisfaction possible … the dynamics were intriguing. (Schelling 2006, pp. 249–250)

This quotation illustrates the two salient characteristics of model-building processes. First, Schelling constructed an epistemic surrogate after failing to find sociological descriptions of the phenomenon that interested him. Initially the surrogate system consisted of symbols on paper, and later of coins on a checkerboard. He was still interested in real phenomena of a certain kind, but now investigated his surrogate models in order to learn about them. Secondly, he modified his models in order to obtain more interesting dynamics. He replaced the symbols on paper with rows of coins because the latter could be handled more quickly and conveniently. He replaced the row of coins with two-dimensional patterns because they allow more variety. However, this set of possible patterns is then narrowed to the checkerboard pattern because it allows for the specification of clear moving rules.

The above quotation also shows that Schelling is not concerned about satisfying any of the stronger model requirements. Although calling the symbols and coins neighbours, and their patterns neighbourhoods, and attributing preferences to them, he makes no effort to justify these labels by pointing out any resemblance to concrete real-world situations, or by citing regularities about the real world. Of course, some very general features—such as the distinguishability of tokens and the spatiality of patterns—resemble features of real-world neighbourhoods. Yet he does

not justify the way in which the model specifies all these features—neighbours' preferences, how neighbours are distinguished, the structure of the neighbourhood, or how discontented neighbours move—with reference to the real world at all. Rather, as the quotation shows, all these specifications are determined by what was available as a formal description that would yield the most interesting and well-defined internal dynamics. Neither similarity, isolation nor conforming to regularity are explicit concerns in Schelling's original paper or in his afterthoughts.

Schelling's case does not seem to be an exception in this respect to economics modelling in general. Modellers often do not argue that their models are linked to the real world, neither through resemblance to some particular system nor through adherence to some law-like regularity. A survey of research articles in economics revealed that almost half of them never make reference to any kind of data.[4] Without such reference, however, neither comparisons nor resemblance claims between models and particular real-world systems can be supported. In particular, without reference to data there is no way of telling whether Giere's theoretical hypothesis is true. Nor is it possible to identify the parts of the real-world system that the model purportedly resembles, and hence it is not possible to judge, as Mäki suggests, the truth of the model concerning those parts.

Further, there is a marked lack of laws and nomic principles in economics, which means that few principles can be used in economic model building. Of the few that exist, the most important is the theory of utility maximisation. Microeconomic models are commonly constructed on a framework derived from that theory, yet there is considerable doubt as to whether it is true or even has any empirical content. What is possibly even more important, modellers do not consider it significant whether or not the theory from which they construct their models is valid. As Samuelson once suggested when discussing the conditions of consumer behaviour implied in expected utility theory: "I wonder how much economic theory would be changed if either of the two conditions above [homogeneity and symmetry] were found to be empirically untrue. I suspect, very little" (Samuelson 1963, p. 117). Such a position implies that economic models cannot be seen as ways of deducing the effects from true principles. Hence they do not capture nature's capacities and workings, and cannot be used to devise Galilean thought experiments, as Cartwright requires.[5]

To conclude, economic modellers 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. In particular, there is a notable lack of explicit attempts to make such references.

[4] These differences were documented for some of the main economic journals during the period from 1972 to 1986. About 45% of all economic research articles analysed mathematical models without making use of or even referring to any form of data, while the respective figures were 18% in political science, 1%in sociology, 0% in chemistry and 12% in physics (Leontief 1982; Morgan 1988). It is my impression that these proportions have not significantly changed in economics (although they may, as part of the methodological aspect of economics imperialism, have increased in the other social sciences).

[5] I should add that Cartwright (2007; this issue) arrives at a similar conclusion. Yet for her this conclusion is bad news for economics—while I would consider it bad news for the 'stronger model requirements' position.

Hence, even if they have a vague idea of what they want the model to resemble, such an idea—one might suspect—is too vague to justify the specific features of model systems.

This lack of reference suggests that many modellers do not heed model properties such as similarity, partial resemblance, or adherence to regularities about the world, and that many models do not satisfy these properties. Approaches to economic modelling such as those of Cartwright, Hausman and Mäki, which stipulate that models should satisfy these properties, or at least that modellers should attempt to satisfy them, thus do not account for such models or the efforts put into constructing them. In the rest of this paper I therefore address the question of whether one can learn from minimal models, irrespective of the satisfaction of stronger, 'world-linking' properties.

5 Learning from Credible Worlds

Sugden (2000; this issue) has put forward the most relevant arguments supporting the possibility of learning from minimal models. He claims that models should offer learning opportunities, insisting that 'model-building has serious intent only if it is ultimately directed towards telling us something about the real world' (Sugden 2000, p. 1). Yet he is critical of both Mäki's isolation account (Sugden 2000, pp. 16–19) and Cartwright's capacity account (Sugden 2000, pp. 20–21; this issue, section 6). Instead, he offers an account of models as *credible* but counterfactual worlds, *paralleling* the real world rather than isolating features of reality.

Sugden's account is similarly minimalist as the one presented here, requiring neither established resemblance nor confirmed nomic principles in model construction. Instead, he argues that we learn from credible model worlds by means of *inductive inference* to the real world.[6] Model users infer from 'a particular hypothesis, which has been shown to be true in the model world, to a general hypothesis, which can be expected to be true in the real world too' (Sugden 2000, p. 19). For example, he argues that Schelling's checkerboard model was intended to support a very general claim: "What Schelling has in mind … [is]: For *all* multi-ethnic cities … strongly segregated neighbourhoods will evolve…" (Sugden 2000, p. 19, my emphasis). We thus infer from the particular instance of a causal mechanism in the imaginary model world to the claim that the same causal forces operate in all situations of the designated category (in this case, US urban residential areas). What could justify such an inference? In other words, what qualities of the model would allow one to increase one's confidence in this general hypothesis? According to Sugden, it is the model's *credibility* that supports the inference.

Sugden proposes that a model is credible only if the situation it depicts is 'possible', 'could be real', or is 'parallel to the real world', in the sense that it

[6] By induction Sugden presumably means the broad category of non-deductive inferences (i.e. all those in which the premises of an argument are believed to support the conclusion but do not entail it), and not the narrower inference from empirically confirmed tokens to other tokens of the same type.

conforms to our experiences and intuitions about the causal forces that operate in the real world.

> … the model world could be real … it describes a state of affairs that is *credible*, given what we know (or think we know) about the general laws governing events in the real world. (Sugden 2000, p. 25)

More specifically, models are credible only if they are logically consistent, and if they 'cohere with common intuitions and experience' (Sugden 2000, p. 26).[7]

The credibility of a model, Sugden argues, contributes to the justification of inference from it to the real world: 'we can have more confidence in them [inferences from models to the real world], the greater the extent to which we can understand the relevant model as a description of how the world *could* be' (Sugden 2000, p. 24).

I have two reservations about this argument. First, I am sceptical about Sugden's claim that the inference goes from claims about a particular model to a general hypothesis about the world. The credibility of a model depends on the intuitions and experiences elicited from *particular* situations.[8] For example, we may judge a model in which agents decide on a certain problem by rule-of thumb heuristics to be credible—while we may at the same time judge it incredible that agents would thus decide on a certain other problem (or even the same problem in a different environment), instead of using a stricter optimisation method. Credibility judgements are highly contingent on the specifics of the situation presented. Further, Sugden claims that when we judge a model to be credible, we say that the causes it depicts also operate in the real world. Yet we do not say that these causes *always* operate in the real world. Sugden, in claiming that one can infer general claims about the world, thus implicitly assumes the stability and universality of causal principles.[9] Without this assumption, credibility judgments only support claims about causes in some particular real-world situation. Model-to-world inferences thus go from particular imaginary situations to *particular* real situations, and not to general claims about the world.

Secondly, even if a model is judged to be credible, and hence 'parallel' to the real world, considerable differences between it and any real-world situation are likely to remain. The question therefore arises why judging the model to be credible will license the application of its causal claims to such a situation. Sugden does not offer a justification, but favours a conventional perspective on inductive practices

[7] Sugden also mentions credibility in a different way, as a quality of the inferences themselves. He writes, 'Since the same effects are found in both real and imaginary cities, it is at least *credible* to suppose that the same causes are responsible' (Sugden 2000, p. 24, my emphasis). Here 'credible' is used in the sense of 'more probable'. This is incompatible with the notion of credibility as depicting a parallel—i.e. counterfactual—reality: the descriptions of counterfactual worlds are necessarily false, and hence cannot be probable. In personal communication with the author, Bob Sugden has suggested that the above use of 'credible' is spurious, and that replacing 'credible' with 'reasonable' or 'defensible' would be a way to avoid possible confusion.

[8] I therefore disagree with the claim that model credibility implies robustness of the results (cf. Kuorikoski and Lehtinen, this issue, section 5).

[9] As further supported by his reference to 'the *general laws* governing events in the real world' (Sugden 2000, p. 25, my emphasis).

(cf. Sugden 2000, footnote 19; cf. also this issue, section 5). Yet, even according to his own account the credibility of the model has to be somehow transferred into confidence in the inference from it to the situation. In other words, we at least have to show that the differences between the two do not give reason to judge as incredible in the real-world situation what was judged as credible in the model. Such an argument requires more than just a credibility judgment of the model: it requires consideration of information about the real-world situation, and comparison to the model world. It seems, then, that credibility alone does not facilitate learning from a minimal model.

My disagreements with Sugden mainly concern *what* we can learn from minimal models, not *whether* we can learn from them at all. I think that his general approach and his emphasis on credibility are sound. In the following I will discuss the notion of credibility in more detail, and then argue that we learn about various possibilities from credible minimal models.

6 Credibility

Credibility is commonly understood as 'the quality or power of inspiring belief' (Merriam-Webster), or 'the quality of being convincing, of that which can be believed' (OED). Examples of credibility judgments regarding models range from finding 'nothing inherently impossible' in a mechanism while remaining unconvinced of its likely reality (to quote an example that Sugden uses, this issue, section 5), to saying that 'it is possible that [the Nash solution agreement] makes sense to people. It is probably one type of argument, from among many, which plays a role in negotiations' (Rubinstein 2001, p. 621). These examples raise two questions: (1) Does judging a model to be credible imply believing in it, or judging it to be true or highly probable? (2) Is credibility in such a judgement attributed to the whole model, or only to certain constituent elements of it?

Regarding the first question I can see two interpretations. One could judge a model to be credible in the sense that one believes it to be true or probably true: the term credibility was sometimes used in this way in the past.[10] In order to apply this notion to models they would have to be translated into propositions or sentences. Given their complex nature this may prove to be difficult, but even if it could be done we face the problem that models are counterfactual situations. A description of them would yield propositions that were false, as they describe worlds that differ from reality. Nevertheless, false statements have low probability and hence would not be considered credible by this account.

One way to avoid this conclusion is to formulate the model descriptions as counterfactual conditionals: *if* the initial and boundary conditions of the model were true, *then* the results would be true. On the one hand, by formulating the descriptive statements as conditionals (with the initial conditions as antecedents) we avoid the conclusion that the model description must be false because some of its initial

[10] For example, Russell (1948) uses 'credibility' in the same way as 'confidence in' or 'degree of belief in' the truth of that proposition.

conditions are false. On the other hand, by formulating them as counterfactual conditionals we avoid the conclusion that because the antecedent conditions are false, the conditional model descriptions must be true (as they would be if they were material conditionals). Avoiding both of these cases would allow the attribution of different probabilities, and hence different degrees of credibility, to model descriptions.

However, this proposal is problematic for at least three reasons. The first of these is practical. Explicating credibility with reference to counterfactuals does not help in the appraisal of economic models, as assessing the truth of counterfactual conditionals is notoriously difficult. Even if the explication were theoretically sound, it would leave us with the formidable task of determining the truth of the *explicans*.

Secondly, and in the light of this practical problem, it seems to me that explicating credibility by means of counterfactuals is like putting the cart before the horse. If one looks more closely at the literature one finds that the very truth of counterfactuals is explicated with reference to something akin to credibility judgments. Most accounts, following Lewis (1973), use some sort of similarity relation between possible worlds. This similarity relation is theoretically basic, but is supposed to be supported by intuitive judgment in combination with knowledge of natural laws. Sugden proposed related arguments, claiming that a model was credible if it cohered with 'the general laws governing events in the real world' (Sugden 2000, p. 25).

Yet there is a conspicuous absence of such laws both in economic papers and in economic seminars presenting models. Modellers commonly start out with the (very wide) utility-maximising framework, and then constrain it by assuming certain functional forms, certain objectives, and so on. They do not usually justify the introduction of these constraints with reference to well-founded empirical regularities of the world.[11] Rather, they justify them as credible or plausible, referring only to their own and to others' intuitions.

Further, a model may be credible even if its dynamics are at odds with established laws or mechanisms. Modellers sometimes study systems that do not—and according to accepted regularities, cannot—exist, such as perpetual motion machines and non-aromatic cyclohexatriene (Weisberg 2007, p. 223). Economists often make similar assumptions, such as infinite consumer sets with infinite lives and immediate consumption (cf. Zamora Bonilla and de Donato, this issue, section 4). These models are not judged incredible on account of these non-law-abiding properties. Thus it seems that credibility does not require adherence to 'general laws', and that models whose dynamics are governed by false fictional principles could still be judged credible.

Others have argued that our commonsense intuitions are generalizations from experience, and that they constitute law-like folk regularities that support credibility judgments (Aydinonat 2007, p. 439). For this argument to go through, credibility judgments would have to consist in comparing the model world with such folk

[11] The exception here is Behavioural Economics, which often refers to experimental results as a way of justifying certain constraining model assumptions.

notions. Yet, first, there is no evidence in research papers or seminars that such comparisons take place, and secondly, there is little evidence that there is a substantial body of folk knowledge in the explicit form of a theory or a set of regularities that would even in principle allow for such a comparison. Instead, folk knowledge of such a kind has resisted any explicit formulation so far. Rather than explicating credibility judgments as a check on whether the model adheres to folk knowledge, it is the folk knowledge that is explicated as the credibility judgments of certain models. Thus, counterfactual judgments in economics cannot be reduced to coherence with either explicit or implicit regularities. What is left for assessing counterfactuals is intuition. Thus it is not that credibility is explicated by means of counterfactuals, it is the truth of counterfactuals that is explicated by means of credibility judgments.

My third concern is with another aspect of the usefulness of the proposed explication. The initial conditions of a counterfactual conditional that describes a model must, at least in principle, be satisfiable in the real world, otherwise the counterfactual conditional account would yield models as not false (and hence capable of positive probability attributions) only in the trivial sense in which a material conditional is true if its antecedent is false. Arguably, however, there are some conditions built into models that are not satisfiable, or their satisfaction cannot be determined. These include 'derivation facilitators', such as the continuity of a distribution, the differentiability of a function (Alexandrova 2006), and 'tractability assumptions' (Hindriks 2006). How would we deal with these assumptions—would they be part of the antecedent, or would they be suppressed? How would we identify them? The difficulties with these questions, I think, point to the difficulties with the whole proposal of linking credibility to truth. I therefore conclude that explicating credibility as truth or resemblance (Mäki, this issue, section 3), realisticness or likelihood (Zamora Bonilla and de Donato, this issue, section 2) leads to profound and, in my view, fatal complications.

Instead, I propose an analogy between the credibility of models and the credibility of fiction. Various authors have linked scientific models to fiction,[12] but here I am focussing on the analogies between the *assessment* of scientific models and literary fictions, which to my knowledge was first suggested by Robert Sugden.

> Credibility in models is, I think, rather like credibility in 'realistic' novels. In a realistic novel, the characters and locations are imaginary, but the author has to convince us that they are credible – that there could be people and places like those in the novel. (Sugden 2000, p. 25)

Works of fiction express fictional propositions. Although these propositions are (commonly) true in the work of fiction in which they are expressed, they are not true: 'Hamlet is a Danish prince' is true-in-*Hamlet*, but it is not true because there is no Hamlet. Further, propositions that are true may be false in fiction. Thus, truth and truth-in-fiction are distinct notions. An important implication of this is that fictional propositions are not supposed to be believed to be true. Fictional narrative does not include real-world referents (in contrast to, say, 'libel' handbills that were used spread

[12] For an overview, see Frigg (2009, p. 6).

lies about real persons), and does not claim to describe the real world (as the documentary genre does, for example,). Rather, it depicts imaginary situations that in important aspects do not resemble the real world: they are not to be believed but are to be imagined. The literary theorist Catherine Gallagher describes this quality as presenting a 'believable story without soliciting belief' (2006, p. 340). In other words, a credible novel has all the features of an account that could well be believed yet because it is imaginary it must not be believed to be true. Instead, the reader is supposed to remain in a state of 'ironic credulity' in which he or she can form a judgment 'not about the story's reality, but about its plausibility' (Gallagher 2006, p. 346).

Imagination commences from fictional descriptions—the text or the spoken word, for example—but goes beyond it. Using description, background beliefs and intuitions, the imaginer creates a fictional world by filling in gaps in the description, adding details, and connecting the discrete accounts to form a continuous, coherent whole. Such an imagination attempt may fail. The elements of a fictional description may not be sufficient for an imaginer to create a coherent and sufficiently complete imaginary world, or they may give rise to incoherence or even contradiction. Depending on the degree of sufficiency and coherence found, the imaginer will judge the description to be plausible or credible. Imagination creates *what* could have been, and assessment of this imagination focuses on *whether* it could have been. The resulting judgment is thus derived from the imaginative activity itself, and not only from consideration of the fictional description.

Models share these two aspects of fiction. First, as I have argued above, they are assessed not according to their truth, but according to their plausibility or credibility. Secondly, judging a model to be credible is a consequence of what scientists do with models: they imagine a world that the model describes, they manipulate that situation in various ways, and they investigate that world's internal coherence and its coherence with our intuitions. Crucially, these intuitions often do not exist independently of the imagined world. Most of us, I suggest, do not hold independent beliefs about how Schelling's token 'agents' behave in a checkerboard-like structure, for example. Instead, vague and rather unspecified intuitions are brought into focus through consideration of such imaginary worlds: It is only when I consider the specifics of the checkerboard model that I judge that something could have been this way. Thus, economic models not only serve as a tool of belief inference, but also *elicit* new beliefs (about something being possible). This role of eliciting beliefs does not depend on the imaginer's believing something to be true or probable. On the contrary, credibility judgments about economics are often elicited solely through consideration of imaginary worlds.

The second question I posed at the beginning of this section concerns the precise object of the credibility judgment—is it the whole model world, or only certain elements of the model? When fiction is judged for its credibility it is notable how many elements of a fictional world do *not* influence the judgment. The exact fictional setting, the properties the characters have and the final outcome of the fictional event are issues that in themselves are irrelevant as far as credibility judgements are concerned. For example, whether the action takes place on the Moon, whether the main character is a demon, or whether in the end he or she attains eternal life are not reasons per se to judge the fictional account incredible.

Rather, it is how the fictional environment, and the fictional characters and their development fit together. For example, Samuel Richardson, the author of *Clarissa*, defended one of his fictional characters against the charge of 'improbability' by giving details from the novel about the specific circumstances in which Lovelace operated (Gallagher 2006, pp. 343–344). His aristocratic background, the over-permissiveness of his mother, the general admiration paid to him by his peers, all this and more supports the plausibility of his actions, which otherwise would seem incomprehensibly vacillating. It is this relation between background information and character development that is the object of credibility judgment, and not the whole fictional setting, from its initial to its final state.

In a similar way, modellers argue for the credibility of their models. All economic micro-models and all micro-founded macro-models depict 'agents' whose behaviour is motivated by some set of beliefs and desires, which in turn are conditional on the agents' perceptions of their environment. These agents 'live' in a fictional environment that the modeller creates. Users of the model judge whether an agent's modelled perceptions are credible *given* the model environment: whether the beliefs and desires attributed to an agent are credible *given* his or her perceptions of the environment and assumed reasoning abilities, or whether the actions are credible *given* his or her beliefs and desires. These conditional credibility judgments are driven by empathy, understanding and intuition. However, they commonly do not extend to the macro level: if the credible interaction of individual agents leads to a surprising or even counterintuitive macro-result, the model is commonly not judged to be incredible. As in a murder mystery, convincing the model user of the credibility of the individual agents' motives, beliefs and actions may convince the user that the initially implausible macro-phenomenon constituted by these individual pieces is also credible. This is precisely the basis for learning from minimal models.

I therefore conclude that correctly judging models to be credible does neither imply that they are true, nor that they resemble the world in certain ways, nor that they adhere to relevant natural laws. Instead, judging models to be credible is a *sui generis* model assessment that is weaker than the conditions discussed in Sect. 4—in particular, this judgment does not require any world-linking properties. Thus we have a way in which to assess minimal models that does not undermine their minimal status. The credibility of a minimal model establishes that it depicts a possible world, a scenario of how the world could be. Such possibility judgments are based on different foundations than judgments concerning how the world is in reality.

7 Learning About Relevant Possible Worlds

Many domains of knowledge contain necessity or impossibility hypotheses. While such hypotheses are sometimes established by stringent scientific means, they are more often entertained in contexts of ignorance about which factors are at work. In such contexts people often maintain general principles—typically about the necessary connection or impossible coexistence of certain factors—that are putatively 'obvious', 'intuitive' or 'common sense'. The set of such principles

pertaining to the domain of economics is sometimes termed 'folk economics', their origin being determined in the untrained human perception of economic phenomena instead of through rigorous scientific research. The principles of folk economics are not necessarily false, but they sometimes contradict current economic theory.[13] Typical examples include claims that higher taxation yields higher state revenues; that free international trade is a zero-sum game between nation states; that immigration hurts national workers; or that a minimum wage improves the lot of the poor.[14] Impossibility hypotheses of this sort constitute an important part in the body of economic knowledge. I will show in this section that we learn from models about the world because consideration of certain models affects our confidence in such impossibility hypotheses.

Necessity and impossibility hypotheses have the logical form $\forall x$: $Px \rightarrow Qx$. In cases in which folk notions are strong people typically attribute relatively high levels of confidence to such hypotheses—and also to the implications, in particular the hypothesis that it cannot be the case that both Px and not Qx. Considering a model may affect these beliefs. Concluding that a model world is credible implies believing that the model presents a possible situation. Such a possible world may exhibit relevant instances of Px that are not also instances of Qx. The person who considers this model to be credible and also has confidence in the impossibility hypothesis thus believes both that $\forall x$:($\neg Px \vee Qx$) and that it is possible that (P & $\neg Q$): such beliefs are inconsistent if the possibility falls within the domain of the quantifier. In order to retain consistency in the belief set, something has to give. If the model cannot be safely rejected as irrelevant, then the person is forced to lower his or her confidence in the impossibility hypothesis. Thus, it seems, we can learn from certain models about the world.

A good example of such a learning effect is Schelling's checkerboard model. Before the models' publication, it seems, many people believed that segregation was necessarily a consequence of explicitly racist preferences. Schelling's model showed that there were plausible settings in which this was not so. By presenting it he thus forced people to change their confidence in the racism hypothesis.[15]

Another example is provided by Schlimm (2009), who discusses cases of learning from the 'mere existence' of cognitive psychology models. He argues that there was widespread belief in the early twentieth century that intelligent behaviour could not be produced without some 'vitalistic' element present in the organism. He then shows that the construction of Hull's psychic machines, Walter's tortoises and Newell and Simon's simulations was directed against this claim. They demonstrated that systems could exist—either in vitro or in silico—which did not contain any 'vitalistic' element, but nevertheless exhibited intelligent behaviour. By 'exhibiting the *existence* of a model of a certain kind', Schlimm concludes, they contradict the

[13] It is sometimes claimed that folk economics has predictable biases, focusing on wealth and its distribution, and neglecting the production and allocation of goods and their efficiency (cf. Rubin 2003).

[14] Economic theory, as well as folk economics, also proposes impossibility hypotheses. Significant examples include the claim that a firm's investment is totally independent of its liquidity position (Modigliani-Miller Theorem), and that the stability of the economy is neutral with respect to the systematic reaction of monetary policy to the business cycle (Rational Expectations Hypothesis).

[15] Tobin's ultra-Keynesian model is another good example (Knuuttila, this issue, section 6).

necessity claim, and in this sense 'scientists can learn something just from the bare existence of models' (Schlimm 2009, p. 19).

If we are to learn from a model, therefore, it must (1) present a relevant possibility that (2) contradicts an impossibility hypothesis that is held with sufficiently high confidence by the potential learners. That the possibility presented is relevant may be supported with reference to natural laws covering this case or to similarity with empirical studies, or as I argued in the previous section, by the credibility of the model.[16]

Yet we do not learn from every credible model. If condition (2) is not met, the model merely shows the possibility of a state that no one believed to be impossible. Such a model would not affect anyone's confidence level, and hence would not have a learning effect. Cases of such epistemically futile models abound in economics. Lind (2007), for example, reviewing recent research on urban economics, finds that most authors claim that 'something can be the case': rent control may have certain consequences, or one can design rules and contracts whereby it may have certain consequences (Lind 2007, p. 7). It therefore seems that the modellers aimed at this learning effect when they built their models. However, in the cases Lind reviews, it is not likely, and the authors could not reasonably have expected, that many people entertained an impossibility hypothesis that their model was able to correct. Therefore, nothing was learned from the models—they fail with respect to their epistemic function.

In cases in which both conditions are met, credible models help in terms of pitching beliefs about possibilities against beliefs about impossibilities: beliefs about credible individual behaviour on the one hand against beliefs about the shape, form and dynamics of aggregate social entities on the other. Minimal models elicit and order these beliefs about individuals, and derive aggregate consequences from them. Such a derivation from a credible individual basis destabilises beliefs about these aggregates, affects our confidence in hypotheses about them, and hence constitutes learning.

8 Conclusion

We learn from minimal models. In particular, consideration of minimal models may lead to a change in our confidence in necessity or impossibility hypotheses. Credible minimal models are sufficient to produce this learning effect—stronger requirements such as resemblance or capacity claims are not necessary. This is an important result, and it shows that economic modelling—even if it does not lead to the further development of testable hypotheses and experiments—is not epistemically futile. Yet it also shows the limits of minimal models: in themselves they do

[16] This irrelevancy problem does not arise in the cases Schlimm discusses. The impossibility claims these models dispute concern the construction of entities that display intelligent behaviour without a 'vitalistic' element. The impossibility hypothesis is about the very possibility of modelling. Thus, the existence of any entity that exhibits such behaviour without incorporating the vitalistic element contradicts the impossibility hypothesis. This feature makes Schlimm's cases special, and prevents their generalisation to other situations.

not support general claims about the world, and they do not support claims about particular real-world situations either. They only play a heuristic role in developing such claims, while the real epistemic support comes from empirical information about the world.

References

Alexandrova, A. (2006). Connecting rational choice models to the real world. *Philosophy of the Social Sciences, 36*(2), 173–192.

Aydınonat, N. E. (2007). Models, conjectures and exploration: An analysis of Schelling's checkerboard model of residential segregation. *Journal of Economic Methodology, 14*(4), 429–454.

Cartwright, N. (1999). Capacities. In J. B. Davis, D. W. Hands, & U. Mäki (Eds.), *The handbook of economic methodology* (pp. 45–48). Cheltenham: Edward Elgar.

Cartwright, N. (2007). The vanity of rigor in economics. In N. Cartwright (Ed.), *Hunting causes and using them: Approaches in philosophy and economics* (pp. 217–35). Cambridge: Cambridge University Press.

Cartwright, N. (2009). If no capacities then no credible worlds. But can models reveal capacities? *Erkenntnis*. doi:10.1007/s10670-008-9136-8.

de Donato, X., & Zamora-Bonilla, J. (2009). Credibility, idealisation, and model building: An inferential approach. *Erkenntnis*. doi:10.1007/s10670-008-9139-5.

Frigg, R. (2009). Models and fiction. *Synthese*. doi:10.1007/s11098-008-9313-2.

Gallagher, C. (2006). The rise of fictionality. In F. Moretti (Ed.), *The novel* (Vol. 2, pp. 336–363). Princeton: Princeton University Press.

Gibbard, A., & Varian, H. R. (1978). Economic models. *The Journal of Philosophy, 75*, 664–677.

Ginits, H. (2000). *Game theory evolving*. Princeton: Princeton University Press.

Giere, R. N. (1988). *Explaining science: A cognitive approach*. Chicago: University of Chicago Press.

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

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy, 21*, 725–740.

Grüne-Yanoff, T., & Schweinzer, P. (2008). The roles of stories in applying game theory. *Journal of Economic Methodology, 15*(2), 131–146.

Hausman, D. (1992). *The inexact and separate science of economics*. Cambridge: Cambridge University Press.

Hindriks, F. A. (2006). Tractability assumptions and the Musgrave-Mäki typology. *Journal of Economic Methodology, 13*(4), 401–423.

Hughes, R. I. G. (1997). Models and representation. *Philosophy of Science, 64*, S325–S336.

Knuuttila, T. (2009). Isolating representations vs. credible constructions? Economic modelling in theory and practice. *Erkenntnis*. doi:10.1007/s10670-008-9137-7.

Kuorikoski, J., & Lehtinen, A. (2009). Incredible worlds, credible results. *Erkenntnis*. doi:10.1007/s10670-008-9140-z.

Leontief, W. (1982). Academic economics. *Science, 217*, 104–107.

Lewis, D. (1973). *Counterfactuals*. Cambridge, Massachusetts: Harvard University Press.

Lind, H. (2007). The model and the story told: An evaluation of mathematical models of rent control. *Regional Science and Urban Economics, 37*(2), 183–198.

Mäki, U. (1994). Isolation, idealization and truth in economics. In B. Hamminga & N. De Marchi (Eds.), *Idealization-VI: Idealization in economics* (pp. 67–68). Poznan studies in the philosophy of the sciences and the humanities, 38. Amsterdam: Rodopoi.

Mäki, U. (2004). Realism and the nature of theory: A lesson from J.H. von Thünen for economists and geographers. *Environment and Planning A, 36*, 1719–1736.

Mäki, U. (2009) . MISSing the world. Models as isolations and credible surrogate systems. *Erkenntnis*. doi:10.1007/s10670-008-9135-9.

Morgan, T. (1988). Theory versus empiricism in academic economics: Update and comparisons. *Journal of Economic Perspectives, 2*, 159–164.

Morgan, M. S. (1999). Learning from Models. In Morgan and Morrison (1999), (pp. 347–388).

Morgan, M. S., & Morrison, M. (Eds.) (1999). Models as mediators: Perspectives on natural and social science. Cambridge: Cambridge University Press.
Rubin, P. H. (2003). Folk economics. *Southern Economic Journal, 70*(1), 157–171.
Rubinstein, A. (2001). A theorist's view of experiments. *European Economic Review, 45*, 615–628.
Russell, B. (1948). *Human knowledge: Its scope and limits*. New York: Simon and Schuster.
Samuelson, P. A. (1963). *Foundations of economic analysis*. London: Oxford University Press (first edition 1947).
Schelling, T. (2006). *Strategies of commitment and other essays*. Cambridge, MA: Harvard University Press.
Schlimm, D. (2009). Learning from the existence of models. On psychic machines, tortoises, and computer simulations. *Synthese*. doi:10.1007/s11229-008-9432-5.
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), 1–31.
Sugden, R. (2009). Credible worlds, capacities and mechanisms. *Erkenntnis*. doi:10.1007/s10670-008-9134-x.
Weisberg, M. (2007). Who is a modeler? *The British Journal for the Philosophy of Science, 58*(2), 207–233.