Erkenn (2009) 70:119–131
DOI 10.1007/s10670-008-9140-z

ORIGINAL ARTICLE

Incredible Worlds, Credible Results

Jaakko Kuorikoski · Aki Lehtinen

Received: 15 April 2008 / Accepted: 25 September 2008 / Published online: 9 January 2009

Abstract Robert Sugden argues that robustness analysis cannot play an epistemic role in grounding model-world relationships because the procedure is only a matter of comparing models with each other. We posit that this argument is based on a view of models as being surrogate systems in too literal a sense. In contrast, the epistemic importance of robustness analysis is easy to explicate if modelling is viewed as extended cognition, as inference from assumptions to conclusions. Robustness analysis is about assessing the reliability of our extended inferences, and when our confidence in these inferences changes, so does our confidence in the results. Furthermore, we argue that Sugden's inductive account relies tacitly on robustness considerations.

1 Introduction

Questions about model-world relationships are questions of epistemology. Many writers treat the epistemology of models as analogous to that of experimentation: one-first builds something or sets something up, then investigates the properties of that constructed thing, and then ponders how the discovered properties of the constructed thing relate to the real world. Reasoning with models is thus essentially learning about surrogate systems, and this surrogative nature distinguishes modelling from other epistemic activities such as "abstract direct representation" (Weisberg 2007; see also Godfrey-Smith 2006). It is then natural to think that the epistemology of modelling should reflect this essential feature: we first learn something about our constructed systems and we then need an additional theory of how we can learn something about the reality by learning about the construct.

J. Kuorikoski (✉) · A. Lehtinen
Department of Social and Moral Philosophy, University of Helsinki, P.O. Box 9,
00014 University of Helsinki, Finland
e-mail: jaakko.kuorikoski@helsinki.fi

Robert Sugden's (2000) seminal paper on "credible worlds" provides a statement of such a "surrogate system" view of models: models are artificial constructs and their epistemic import is based on inductive extrapolation from these artificial worlds to the real world.

The epistemic foundation of models according to this view is thus based on an inductive leap, which is similar to that of extrapolating from one population to another: we first build a population of imaginary but credible worlds, investigate their salient features, and then make a similarity-based inductive leap by claiming that the real world also shares these salient features.

The purpose of this article is to present an alternative perspective on modelling that helps in making epistemic sense of the relationships between models and the world. Our argument is that, from the epistemic point of view, modelling is essentially inference from assumptions to conclusions conducted by an extended cognitive system (cf. Giere 2002a, b). Our viewpoint has some obvious affinities with Suárez' (2003, 2004) inferential account of scientific representation (see also Contessa 2007), but our main intention is not to provide a theory of scientific representation. While we agree with Suárez' main arguments against similarity and isomorphism, we also agree with those who have pointed out that these arguments against dyadic representation apply to representation in general rather than just scientific representation (Brandom 1994; Callender and Cohen 2006). We are also sympathetic to Knuuttila's (2005, 2009) productive view in stressing that models are artefacts used to produce claims, and in downplaying the explanatory primacy of any representational relationship between the world and a model as a whole in accounting for the epistemic value of models.

We do not claim that the surrogate-system view is wrong. We advocate our ontologically deflationist perspective in order to guard against mistakes that may arise from taking this view too literally. The perspective of modelling as extended inference constrains and complements the surrogate-system view. First, viewing modelling as inference constrains its epistemic reach into being the same as that of argumentation: a model does not contain any more information than that which is already present in the assumptions. Our aim is to dispel the impression that there is a special *philosophical* puzzle of how we can learn about the world by simply looking at our models. Secondly, viewing modelling as inference from assumptions to conclusions implies that, in principle, all epistemic questions about modelling can be conceived as concerning either the reliability of the assumptions or the reliability of the inferences made from them. The first is a matter of whether the assumptions are true (or perhaps close to being true) of, or applicable to, some specific real system. The second is a matter of whether the way in which conclusions are derived from these assumptions may lead to false conclusions even when the assumptions are (roughly) true. The reliability of inferences concerns conclusions that are about some real target system. The corresponding within-model inferences are usually deductive and thus maximally secure.

Evaluating the reliability of assumptions may involve various epistemic activities such as formulating intuitive judgments concerning their truthlikeness, testing the assumptions empirically, and so forth. If all the assumptions are true, and the modeller makes valid inferences from them, trivially, the conclusions are

empirically supported. In this case, a within-model inference is simply a model-world inference. The epistemic problem in modelling arises from the fact that models always include false assumptions, and because of this, even though the derivation within the model is usually deductively valid, we do not know whether our model-based inferences reliably lead to true conclusions. Even though modellers sometimes need to make judgments concerning the structural similarity of their model and a target system, there is no need to think of models as abstract objects and thus no special epistemic puzzle of linking abstract or constructed objects to reality. The model's structure ultimately derives from the assumptions and the way in which they are put together, or to put it differently, the structure *is* one of the model's assumptions.

We will argue that robustness analysis is essentially a means of learning about the reliability of our model-based inferences from assumptions to conclusions. Our perspective may thus account for the epistemological significance of robustness analysis and thereby complements the surrogate-system view. In contrast, according to the surrogate-system point of view, robustness analysis is only a matter of comparison between constructed worlds, and cannot therefore be relevant to the model-world relationship. Sugden explicitly makes this argument in his Credible Worlds paper: robustness analysis cannot take us outside the world of models and therefore cannot be relevant to the inductive leap from models to the world.

2 Models as Surrogate Systems

There seems to be an analogy in the *epistemic dynamics* of models and experiments. It is most conspicuous in the case of simulation models: we build a surrogate system, investigate it, and then think of how to apply or relate the findings to the real world. Uskali Mäki (2005) and Mary Morgan (2003) have taken the analogy between models and experiments further by arguing that constructing a surrogate system and setting up an experiment also have certain logical similarities. According to Mäki (esp. 1992, 1994), modelling and experimentation are both attempts at *isolating* the causally relevant factors. In the case of models such isolation is achieved by making more or less unrealistic assumptions (theoretical isolation), and in the case of experiments it is achieved causally through experimental controls (material isolation).

Models are also claimed to be autonomous from theoretical presuppositions (Morgan and Morrison 1999). Their autonomy derives from the fact that obtaining tractable models from a theory always involves making various auxiliary assumptions. That is what modelling is all about: changing, modifying, simplifying and complexifying such auxiliary assumptions, often in a more or less ad hoc manner. Thus even theoretical models, let alone phenomenological or data models, are autonomous in the sense of involving assumptions not derivable from the theory. One possible way of understanding the notion of model autonomy is to say that the underlying theory restricts the results of modelling very little: the results inevitably depend on the auxiliary assumptions. In economics, for example, utility maximisation does not imply all that much in itself, as Kenneth Arrow (1986) has argued.

Thinking of models in terms of autonomous surrogate systems, and reflecting on the apparent similarities between them and real-world experiments, would appear to lead naturally to the idea that the epistemic question concerning the model-world relationship is to be analysed in terms of a model that has already been constructed, an autonomous self-standing construct, and that answering this question requires a special account or a theory. Existing accounts have been based on considerations of similarity (Giere 1988), isomorphism (Van Fraassen 1980; da Costa and French 2003) or simple induction (Sugden 2000). Any answer to this epistemic question should be constrained by the general epistemological position adopted, and we take that position to be level-headed naturalistic empiricism: we can only find out about the world from our experience of the world. For there to be such experience, a suitable causal connection between the cognitive agent and the world is necessary. We do not make any additional arguments for empiricism here because we take it to be the default position in the philosophy of science. However, if empiricism is true, the question of how we can learn something new about the real world merely by studying our models becomes acute. Viewing models as surrogate systems seems to suggest that we are supposed to learn something about the real world by experimenting with and making observations about imaginary or abstract objects. However, experimentation on an imagined or an abstract construct is not the same thing as real experimentation, and finding out that a model has a certain property is not the same thing as making an observation about the target phenomenon. In neither case is there any direct causal contact between the modeller and the modelled system.

3 Modelling as Extended Inference

The question of how to reconcile naturalistic empiricism with the apparent epistemic value of models is the same as that which John Norton (2004a, b) asks about thought experiments. In our view, the correct answer is also the same: the epistemic reach of modelling is precisely the same as that of argumentation. Argumentation here means, roughly, using formal syntactic rules to derive contentful expressions from other contentful expressions in a truth- or probability-preserving manner. What sets modelling apart from pure thought experimentation is that in the former the inferences from assumptions to conclusions are conducted not entirely in the head of the modeller or only in natural language, but rather with the help of external inferential aids such as diagrams, mathematical formulas and computer programs.[1] In de Donato Rodríguez and Zamora Bonilla (2009) words, models function as inferential prostheses. What is doing the cognitive work in modelling is not the individual, but the individual-model pair. Modelling is essentially inference from assumptions to conclusions *conducted by an extended cognitive system* (cf. Giere 2002a, b).

[1] As one referee pointed out, models could be seen as thought experiments of the extended cognitive system.

In our view, although modelling necessarily involves abstracting, models in themselves are not abstract entities. Abstraction is an activity performed by a cognitive agent, but the end result of that activity, the abstraction of something, need not in itself be an abstract entity. Instead, it is (often) a material thing used to represent something. We take models to be things made out of concrete and public representations, such as written systems of equations, diagrams, material components (for scale or analogue models), or computer programs actually implemented in hardware. Abstract objects, insofar as they can be said to exist in the first place, are non-spatiotemporal and causally inert things, and therefore cannot engage in causal relations with the world or the subject. This is why we think that abstract objects cannot play an ineliminable role in a naturalistic account of the epistemology of modelling.

It is usually the inferential rather than the material (or ontological) properties of these abstractions that are epistemically important for the modeller. Although the material means of a representation often do matter in subtle ways for what inferences can be made with it,[2] the aim in modelling is to minimise or control for these influences: if a conclusion derived from a model is found to be a consequence of a particular feature of a material representation lacking an intended interpretation, the conclusion is deemed to be an artefact without much epistemic value. Therefore, it often makes perfect sense to further abstract from multiple individual representations to their common inferential properties and then label these common inferential properties as "the" model itself. These inferential properties are, of course, not intrinsic to the representations, but rather depend on the context in which they are used.[3] There is thus no need to abandon the distinction between "the model" and its various descriptions (cf. Mäki 2009). For example, many kinds of public representations facilitate similar kinds of inferences, from spring constants and amplitudes to total energy, and this makes all of these representations models of *the* harmonic oscillator. Such abstractions are often extremely useful in co-ordinating cognitive labour. By referring to them we refer only to a set of inferences and can therefore disregard the material things that enable us to make these inferences in practice. The material form these representations may take is usually not relevant to the epistemic problems at hand: whether a differential equation was solved on a piece of paper, on a blackboard, or in a computer is not usually relevant to whether or not it was solved correctly. This is why it is natural to think that the "identity" of the model of the harmonic oscillator resides precisely in these

[2] Much of the recent philosophical literature on models underscores this point. For example, Marion Vorms (2008) uses the example of the harmonic oscillator (simple pendulum) to show how the material means or *the format* of representation may matter in subtle ways to what kind of inferences can be made with it.

[3] Whether a public and a material thing can be used to make inferences about something else naturally depends on the intrinsic properties of the thing in question. Thus the contextual nature of representation does not mean that it is completely arbitrary whether something can be a representation of a particular system. We believe that this is the intuition behind the idea that there has to be a substantial account of representation that explains the epistemic properties of models (see e.g. Contessa 2007). However, noting that the intrinsic properties of things matter to what can be done with them does not yet imply that it is the concept of representation that accounts for the inferential properties of a model, rather than the other way round (cf. Brandom 1994).

common inferential properties of the various material representations, i.e. in the abstract object. Nevertheless, we should resist reifying the abstractions as abstract objects in themselves.

Neither adopting a naturalistic empiricist viewpoint nor denying the causal efficacy of abstract objects should be very contentious. Yet, framing the epistemic situation in this way undermines the notions that the epistemic question of how we can learn from models is only asked after the properties of the abstract object have been investigated, and that there should be a special answer or theory accounting for it (e.g. extrapolation or simple induction from a set of credible worlds to the real world). If modelling is inference, then making valid inferences from empirically supported assumptions would automatically give us empirically supported conclusions.[4] The inductive gap between the model and the world arises from the fact that all of the assumptions are never true and the inference becomes unreliable. We will substantiate this claim further when we discuss robustness analysis in the next section.

Modelling is clearly distinct from ordinary inference and argumentation in that we seem to find genuinely new things by manipulating or investigating an artificial construct. This analogy between the epistemic dynamics of modelling and experimentation can be misleading, however. The discovery of novel information is often experienced as a similarity between modelling and experimentation. The sense of novelty is a result of the essential use of external inferential aids in modelling. Using mathematics, diagrammatic reasoning carried out with pen and paper, or computer simulation involves manipulations of representations external to the mind of the human subject, and he or she may not experience this manipulation as inference-making, i.e. as phenomenologically similar to thinking. What the human subject experiences is more akin to experimentation with an artefactual, abstract or imaginary system. The modeller manipulates graphs or mathematical equations—something external to his or her mind—and then finds something new about the abstract object that is represented by the equations or graphs. Yet, from the perspective of the whole extended cognitive system consisting of the modeller and the external representations (the model), there is no experimentation, only inference. The only "epistemic access" (cf. Mäki 2009) that the extended cognitive system has to the target system is via the original causal connection that was required for the formulation of the substantial empirical assumptions from which the inferences are made. The things that are found out are new only in the sense that the conclusions were not transparent to the unaided reasoning powers of the modeller.

Giere (2002a) uses a simple example of how even elementary arithmetical operations essentially require the manipulation of external representations. Multiplying three-digit numbers in our head is beyond the cognitive capacities of most of us. By using pen and paper in the way we learned in elementary school, we can break this arithmetic operation down into a series of single-digit multiplications and additions. The human being performing this task is only making single-digit

[4] This does not mean that the model outcome would be empirically supported in the sense that it should straightforwardly agree with the observations. Modelling results are often claims about tendencies or capacities of the modelled systems, and the manifestations of these tendencies can be blocked by factors not included in the models.

inferences, but the extended cognitive system consisting of the human, the pen and the paper is making a three-digit inference. In economics, the use of diagrammatic reasoning offers a vivid example of how manipulating external representations such as supply and demand curves, budget constraints, and indifference curves facilitates inference from assumptions to conclusions. Manipulating mathematical symbols with pen and paper according to formal syntactic rules is a similar activity: it is inference from assumptions to conclusions, which is now encoded in a mathematical formalism.

We do not deny the heuristic value of viewing models as surrogate systems. Such a view captures an important distinction between modelling and other forms of scientific representation and reasoning. However, we claim that it is prone to leading to two kinds of mistakes. First, such a view may tempt one to overestimate the epistemic reach of modelling. Austere empiricism with respect to the epistemology of modelling helps us clarify what kinds of things we may learn through modelling and, crucially, what cannot be achieved. For example, the results of true experimentation may depend on the causal properties of the constituent parts of the investigated system that were unknown when the experiment was designed, whereas the model constituents cannot, by definition, have any other properties than those postulated. In short, the epistemic reach of modelling is the same as that of argumentation.

Secondly, learning about the properties of our inferential aids may be epistemologically relevant, but the epistemic dynamics of such learning need not be similar to learning about the properties of real systems. Viewing models as surrogate systems may thus mistakenly lead us to classify some of the things we do with them as having little or no epistemic relevance. We use derivational robustness analysis,[5] i.e. the practice of deriving the same result using different modelling assumptions, as an example of an epistemological strategy with respect to which this latter mistake can be made.

4 Robustness Analysis and the Reliability of Inferences

Part of the attractiveness of the surrogate-system view of models may derive from the fact that it seems to explain why modellers are often interested only in the properties of the models rather than in the relationship between the model and the world: models are often taken to be interesting systems in their own right. Mäki (2009) laments that merely studying the internal logic of models is not compatible with scientific realism, since questions of truth seem to be eschewed. This accusation prompts the question of whether we can obtain knowledge about the material world by merely investigating models even if we wish to avoid commitment to some kind of problematic rationalism or other kinds of epistemic magic.

When we are primarily interested in the properties of models, we are interested in the properties of our inferential aids. When our confidence about our inferences

[5] See Woodward (2006) for an account of different types of robustness, and Wimsatt (1981) or Weisberg (2006) for an account of its epistemic importance.

changes due to new knowledge concerning the properties of the inference apparatus, our confidence about the conclusions derived using the apparatus also changes. This is how merely learning about models may legitimately change our beliefs about the world. Viewing modelling as extended cognition therefore explains why and how our beliefs about the world may change when we learn more about our models, and shows that this can be done in a way that is consistent with empiricism.

This is why, contrary to Sugden, we think that *derivational robustness analysis* may have epistemic import even though it is a mere comparison between models. All modelling, or at least all theoretical modelling, involves false assumptions, and is thus unreliable as an inferential aid. By unreliability we mean the following: the modeller knows that he or she has to make unrealistic assumptions in constructing the model, but not whether their falsity (in the sense of not being nothing-but-true and the-whole-truth) undermines the credibility of the results derived from it.

Theoretical modelling usually involves roughly two kinds of assumptions: substantive and auxiliary. Substantive assumptions concern aspects of the model's central causal mechanism about which one endeavours to make important claims. They are usually assumptions that, it is hoped, have some degree of empirical merit, i.e. they are thought to be more or less true of the systems on which it is hoped that the model will shed some light. The set of target systems need not be fully specified or even suggested in advance, and the stories that often accompany models could be seen as selling points for their inferential abilities (cf. Sugden 2009). Auxiliary assumptions (tractability assumptions and derivation facilitators, for example) are required for making inferences from these substantive assumptions to conclusions feasible (Musgrave 1981; Mäki 2000; Alexandrova 2006; Hindriks 2006). Different auxiliary assumptions create different kinds of distortions and biases in our inferences. By errors and biases, we do not mean logical or mathematical mistakes in inferences, but rather false consequences that the use of false auxiliary assumptions may lead us to draw about the target phenomenon. For example, Nancy Cartwright is concerned that auxiliary assumptions introduced through the very structure of economic models might create irremediable errors in them that in her words "overconstrain" the results (Cartwright 2009). Given that making at least some unrealistic assumptions is unavoidable, these errors and biases are also unavoidable, and the best epistemic modelling strategy is to accept their inevitability and to try to control for their effects. Modelling practice must thus allow for systematically examining the different roles assumptions play, and thus for at least locating the various errors.

Derivational robustness analysis is the procedure for testing whether a modelling result is a consequence of the substantive assumptions or an artefact of the errors and biases introduced by the auxiliary assumptions. It is carried out by deriving a result from multiple models that share the same substantive assumptions but have different auxiliary assumptions. The main functions of derivational robustness analysis are to root out errors and to provide information about the relative importance of the assumptions with respect to the result of interest (Kuorikoski et al. 2007). By controlling for possible errors in our inferences, robustness analysis makes our conclusions more secure. It could therefore increase (or decrease) our confidence in the modelling results and change our beliefs about the world, although it is, strictly

speaking, only a matter of comparing similar models. Robustness analysis is thus an important part of a thoroughly empiricist epistemology of modelling.

We are now in a position to clarify our claim that if the reliability of assumptions and inferences has been successfully evaluated, there is no further puzzle concerning how the model as an abstract object relates to reality. If the substantive assumptions are empirically well-supported, and if the modeller makes truth-preserving inferences from them, the conclusions are empirically supported. Insofar as the substantive assumptions are realistic and the inference from them is reliable, they also carry their epistemic weight into the results. The role of robustness analysis is to show that the conclusions are not an artefact of the auxiliary assumptions, but rather derive from the substantive assumptions. Even though deriving a result with a different set of auxiliary assumptions usually involves deductive inference, the epistemic importance of robustness analysis is based on the fact that it is not possible to provide a complete list of all logically possible alternative auxiliary assumptions, and is in this sense inductive. Furthermore, modelling results are seldom completely robust, even with respect to the auxiliary assumptions that can be specified, and model-based inferences are therefore less than foolproof. Hence, an inductive gap remains between the substantive assumptions (not "the model") and the world. Whether or not there are further substantial but yet general or even a priori constraints that could be imposed on this inductive leap is an open question.

If the substantial assumptions are not realistic, no amount of robustness analysis suffices to change our views about which results of the model could also be taken to hold in the real world. Robustness analysis is thus useless if all assumptions are unrealistic, and its epistemic relevance rides on there being at least some realistic assumptions. However, the fact that the epistemic status of even substantive assumptions is often unclear only goes to show that robustness analysis is fallible, as all forms of inductive inference are. This does not change the fact that it is the truthlikeness of the substantial assumptions that ultimately carries the epistemic weight in a model.

The logic of investigating the properties of inferential aids need not be, and often is not, similar to that of material experimentation. Robustness analysis is a case in point because it cannot always be considered a theoretical counterpart to causal isolation or de-isolation. Material isolation works by eliminating the effects of disturbances, while derivational robustness analysis works by controlling for errors induced by auxiliary assumptions rather than by eliminating them. This difference has a number of consequences.

Even though it is possible to control for the effects of disturbances, the way in which this happens differs from controlling for errors that are induced by auxiliary assumptions. The crucial difference is that if we know how to control for a disturbance in an experiment, we know how the disturbance affects the phenomenon under scrutiny. In contrast, the auxiliary assumptions are often so unrealistic that it is misleading to think of them as possible causal factors, as the isolation account seems to suggest. It is often simply impossible to define a metric for the truthlikeness of auxiliary assumptions. The epistemic goal of derivational robustness analysis may be served just as well by replacing unrealistic assumptions with equally unrealistic assumptions as with more realistic ones. The crucial question is

not the truthlikeness of the alternative auxiliary assumptions, but rather their *independence*. Independence in this context means, roughly, that the alternatives are false in different ways, and that they are therefore unlikely to create similar biases in reasoning. By deriving the same result using a number of independent sets of auxiliary assumptions, we reduce the possibility that the result is primarily a consequence of an error created by a particular auxiliary assumption.

Acknowledging the importance of derivational robustness helps us to make sense of the idea that models are often best evaluated as whole sets of similar inferential frameworks rather than as isolated and self-standing entities (cf. Aydinonat 2008, pp. 144–166). The question of what exactly determines the identity of a model becomes less important when we realise that evaluation is conducted on the level of sets of models on the one hand, and individual assumptions on the other.

5 Similarity, Credibility and Robustness

Let us finally take a closer look at Sugden's account of models as credible worlds. He argues that the relationship between models and the world can be evaluated in terms of similarity: the credibility of abstract models and the consequent inductive inference are somehow established according to similarity judgments. As Sugden correctly points out, in economics at least, the credibility of models is often argued for by way of providing an empirical illustration or a "story" about how the mechanism at work in the model could be at work in the real world (see also Morgan 2001). We could complain about the fact that he has not really given an account of how similarity judgments are supposed to do their job, but given that others have not come up with much more content for such similarity claims, we will not concentrate on that here.

Similarity alone is usually not sufficient for establishing credibility, and stories with which the credibility of models is buttressed often make essential reference to the robustness of the model—as indeed they do at least in Sugden's own examples. Our argument is thus that Sugden's intuitive notion of similarity presupposes at least an implicit reference to robustness considerations.

Consider now whether it would be sufficient to use mere similarity, without at least an implicit reference to robustness, to establish the credibility of a model. Sugden's examples include Akerlof's (1970) lemons model and Schelling's (1978) checkerboard model. Let us take the latter first. As Sugden notes, the checkerboard representation of racial segregation does not derive from a series of isolation operations that start from a real city. It would be better to say that the checkerboard is the result of constructing a model, and that if the finished model can be taken to isolate some particular causal factors, the isolation operation must be conceived to be an aspect of the whole process of constructing the model.[6] In other words, once Schelling had come up with the checkerboard structure, he was able to isolate

[6] Uskali Mäki has argued (e.g., in 2009) that Sugden's claim about the constructive nature of modelling is implicitly based on isolation: when we have constructed the model, we have already made the necessary isolations.

something that seemed crucial to racial segregation: the idea that segregation could be analysed in terms of individual localisation decisions in a two-dimensional spatial grid. However, the checkerboard structure is in crucial respects extremely dissimilar to real cities: people in real cities often live in some sort of clusters, on top of each other, for example. Without qualification the checkerboard is surely an incredible world. This trivial-sounding observation simply reminds us that similarity comparisons are sensible only with respect to particular aspects of the things compared, and only against a given background context. Checkerboards and real cities are, of course, similar in that they can both be depicted in a two-dimensional space in the first place, and this particular dimension of similarity has something to do with the phenomenon that is being investigated. Surely, however, this similarity alone would not have convinced anybody about the credibility of the checkerboard structure in accounting for racial segregation if whether or not the obvious dissimilarities mattered for the result of this model were an open question. Schelling realised this and claimed (although without actually proving) that the actual geometric shape (two-dimensional or three-dimensional, a grid or a torus, for example) and the initial spatial configuration of individuals on the grid did not matter. Subsequent developments have partly vindicated this robustness claim. At the risk of being speculative, we feel confident in claiming that the checkerboard model would not have become so famous had its credibility not received support from other scholars who showed that it was robust with respect to most of these other assumptions.[7]

What about Akerlof's lemons model, then? The importance of robustness in creating credibility is admittedly less evident than in Schelling's model, but Akerlof's empirical illustrations of the lemons principle do establish the idea that insofar as this principle is at work at all, its consequences will be similar in widely differing circumstances: informational asymmetry results in a reduction of the volume of trade and a deterioration in the average quality of goods. The model is indeed similar to the real world in that it is relatively easy to recognise the fact of asymmetric information in the various settings that Akerlof presents. When he asks us to consider the idea that there are four kinds of cars (new and old, good and bad), he is implicitly referring to a robustness consideration: we think that making the more realistic assumption that cars can be arranged on a continuum with respect to age and quality would not really affect the consequences of incomplete information.

Nothing is similar to something else *tout court*. Meaningful comparisons of similarity can only be made with respect to specific features of the things compared and only against some background context. Similarity confers credibility on a model only when it is the important parts of the model that are, to some degree, similar to the modelled systems. Therefore, even from Sugden's own point of view, robustness of these important parts of the model with respect to auxiliary assumptions (which may or may not be similar to the real world) has to be ascertained before the similarity comparison can do the epistemic work it is supposed to do.

[7] See Aydinonat (2007, 2008) for a review of Schelling's model and for more ways in which it is dissimilar to real cities.

6 Conclusions

It is generally accepted that indirect representation and surrogative reasoning are the cornerstones of the epistemic strategy of model-based science. In this article we have stressed that although this widely shared view is correct, modelling as a cognitive activity is nothing more than inference from assumptions to conclusions conducted by an extended cognitive system, i.e. argumentation with the help of external reasoning aids. This additional perspective helps to dispel some epistemic puzzles that might arise from taking the surrogative and semi-experimental phenomenology of modelling too far.

Our perspective also helps to illustrate how merely looking at models may justifiably change our beliefs about the world. When we learn more about the reliability of our inferences, the reliability attributed to our conclusion should also change. By reliability we mean the security of our inferences against the distorting effects of the inevitable falsities in modelling assumptions. Derivational robustness analysis is a way of assessing the reliability of our conclusions by checking whether they follow from the substantial assumptions through the use of different and independent sets of false auxiliary assumptions. It is a way of seeing whether we can derive credible results from a set of incredible worlds.

References

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics, 84*(3), 488–500.

Alexandrova, A. (2006). Connecting economic models to the real world: Game theory and the FCC spectrum auctions. *Philosophy of the Social Sciences, 36*(2), 173–192.

Arrow, K. J. (1986). Rationality of self and others in an economic system. *Journal of Business, 59*(4), 385–399.

Aydinonat, 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.

Aydinonat, N. E. (2008). *The invisible hand in economics: How economists explain unintended social consequences*. London: Routledge.

Brandom, R. B. (1994). *Making it explicit: Reasoning, representing and discursive commitment*. Cambridge, MA: Harvard University Press.

Callender, C., & Cohen, J. (2006). There is no special problem about scientific representation. *Theoria, 21*(55), 7–25.

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

Contessa, G. (2007). Scientific representation, interpretation, and surrogative reasoning. *Philosophy of Science, 74*(1), 48–68.

da Costa, N. C. A., & French, S. (2003). *Science and partial truth: A unitary approach to models and scientific reasoning*. New York: Oxford University Press.

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

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

Giere, R. (2002a). Models as parts of distributed cognitive systems. In N. J. Nersessian & L. Magnani (Eds.), *Model based reasoning: Science, technology, values* (pp. 227–241). New York: Kluwer/Plenum.

Giere, R. (2002b). Scientific cognition as distributed cognition. In P. Carruthers, S. Stich, & M. Siegal (Eds.), *The cognitive basis of science* (pp. 285–299). Cambridge, UK: Cambridge University Press.

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy, 21*, 725–740.
Hindriks, F. A. (2006). Tractability assumptions and the Musgrave-Mäki typology. *Journal of Economic Methodology, 13*(4), 401–423.
Knuuttila, T. (2005). Models, representation, and mediation. *Philosophy of Science, 72*(5), 1260–1271.
Knuuttila, T. (2009). Isolating representations vs. credible constructions? Economic modelling in theory and practice. *Erkenntnis*, this issue. doi:10.1007/s10670-008-9137-7.
Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2007). Economics as robustness analysis. Homepage of PhilSci Archive: http://philsci-archive.pitt.edu/archive/00003550/econrobu.pdf.
Mäki, U. (1992). On the method of isolation in economics. In C. Dilworth (Ed.), *Intelligibility in science* (pp. 319–354). Atlanta and Amsterdam: Rodopi.
Mäki, U. (1994). Isolation, idealization and truth in economics. In B. Hamminga & N.·B. De Marchi (Eds.), *Idealization VI: Idealization in economics* (pp. 147–168). Amsterdam and Atlanta: Rodopi.
Mäki, U. (2000). Kinds of assumptions and their truth: Shaking an untwisted F-twist. *Kyklos, 53*(3), 317–335.
Mäki, U. (2005). Models are experiments, experiments are models. *Journal of Economic Methodology, 12*(2), 303–315.
Mäki, U. (2009). MISSing the world. Models as isolations and credible surrogate systems. *Erkenntnis*, this issue. doi:10.1007/s10670-008-9135-9.
Morgan, M. S. (2001). Models, stories and the economic world. *Journal of Economic Methodology, 8*(3), 361–397.
Morgan, M. S. (2003). Experiments without material intervention: Model experiments, virtual experiments and virtually experiments. In H. Radder (Ed.), *The philosophy of scientific experimentation* (pp. 236–254). Pittsburgh: University of Pittsburgh Press.
Morgan, M. S., & Morrison, M. (1999). *Models as mediators: Perspectives on natural and social science*. Cambridge: Cambridge University Press.
Musgrave, A. (1981). "Unreal assumptions" in economic theory: The F-twist untwisted. *Kyklos, 34*(3), 377–387.
Norton, J. D. (2004a). Why thought experiments do not transcend empiricism. In C. Hitchcock (Ed.), *Contemporary debates in philosophy of science* (pp. 44–66). Oxford, UK: Blackwell Publishing.
Norton, J. D. (2004b). On thought experiments: Is there more to the argument? *Philosophy of Science, 71*(5), 1139–1151.
Schelling, T. C. (1978). *Micromotives and macrobehavior* (1st ed.). New York: Norton.
Suárez, M. (2003). Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science, 17*(3), 225–244.
Suárez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science, 71*(5), 767–779.
Sugden, R. (2000). Credible worlds: The status of theoretical models in economics. *Journal of Economic Methodology, 7*(1), 169–201.
Sugden, R. (2009). Credible worlds, capacities and mechanisms. *Erkenntnis*, this issue. doi:10.1007/s10670-008-9134-x.
Van Fraassen, B. C. (1980). *The scientific image*. Oxford: Clarendon Press.
Vorms, M. (2008). Models and formats of representation. Homepage of PhilSci Archive: http://philsci-archive.pitt.edu/archive/00003900/01/models_and_formats_of_representation.pdf.
Weisberg, M. (2006). Robustness analysis. *Philosophy of Science, 73*(5), 730–742.
Weisberg, M. (2007). Who is a modeler? *British Journal for the Philosophy of Science, 58*(2), 207–233.
Wimsatt, W. C. (1981). Robustness, reliability and overdetermination. In M. B. Brewer & B. E. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 124–163). San Francisco: Jossey-Bass.
Woodward, J. (2006). Some varieties of robustness. *Journal of Economic Methodology, 13*(2), 219–240.