**How fictional accounts can explain**

**[A comment on ‘The explanation paradox’ by Julian Reiss]**

**Robert Sugden**

School of Economics

University of East Anglia

Norwich NR4 7TJ

United Kingdom

[r.sugden@uea.ac.uk](mailto:r.sugden@uea.ac.uk)

25 June 2012

*Keywords*: fiction, explanation, explanation paradox, credible world

*Abstract*: In this note, I comment on Reiss’s paper ‘The explanation paradox’. I argue in support of two of the propositions that make up the paradox (that economic models are false, and that they are explanatory) but challenge the third proposition, that only true accounts can explain. I defend the ‘credible worlds’ account of models as fictions that are explanatory by virtue of similarity relations with real-world phenomena. I argue that Reiss’s objections to the role of subjective similarity judgements in explanation illegitimately presuppose the existence of rational criteria by which science can certify the objective validity of its explanations.

*Acknowledgement* In writing this paper I have benefitted from discussions with Stefan Heidl.

Julian Reiss’s paper is unusual in presenting arguments in support of a set of mutually inconsistent propositions. Together, the three propositions in this set constitute the ‘explanation paradox’ (Reiss 2012). Clearly, at least one of these propositions, and therefore at least some of Reiss’s arguments, must be rejected. I shall be mainly concerned with the third proposition: ‘Only true accounts can explain’ (p. 49). In previous papers I have argued that theoretical models in economics can be explanatory while also being, in a certain sense, fictional (Sugden 2000, 2009, 2011a). Reiss discusses and rejects this ‘credible worlds’ account of models. I shall defend it against his criticisms.

Reiss considers a class of theoretical models in economics, exemplified by Harold Hotelling’s (1929) spatial model of competition, interpreted as a putative explanation of a ‘principle of minimum differentiation’. Reiss claims, surely correctly, that one of the aims of economics is ‘to give genuinely explanatory accounts of economic phenomena’ (p. 44). Hotelling’s model is an example of a mode of explanation widely used in economics. It is generally considered by economists to be explanatory; and it *feels* explanatory (pp. 48–49). But, according to Reiss, the model is false. Thus, if economists (and Reiss’s feelings) are right, false accounts can explain, contrary to Reiss’s third proposition. So why can’t we just reject that proposition? Referring his readers to the literature of philosophy of science, Reiss reports that ‘the causal account of explanation is widely regarded as successful’, and claims that, as a matter of conceptual necessity, ‘causal explanations cannot be successful unless they are true’ (pp. 43, 45). The upshot of this is that Reiss’s paradox is essentially an opposition between what counts as an explanation within the practice of economics, and what (we are told) most philosophers of science think *ought* to count as an explanation. One might add that the causal account of explanation, as presented by Reiss, accords with common sense: in everyday discourse, explanations of events do usually cite their supposed causes, and most people would find it hard to believe that a false statement could explain anything at all. My account of models as credible worlds defends the practice of economics against both philosophy and common sense.

Before explaining why I think Reiss’s arguments against that account cannot be sustained, I will offer some support for his first two propositions – that economic models are false, and that they are nevertheless explanatory. This may help to explain why I think the credible-worlds account of models is necessary.

As Reiss points out, some methodologists argue, contrary to his first proposition, that an apparently unrealistic model can be true ‘in the abstract’. That is, such a model truly describes a particular mechanism that operates in the real world, but it does so by isolating that mechanism from others than typically operate alongside it (Mäki 1992, 2009; Cartwright 1998, 2009). If this were read only as a claim that a modeller makes about his or her model, it would be unexceptionable. For example, Hotelling’s model undoubtedly describes a mechanism that induces non-differentiation between the locations of the duopolists in his model world. Hotellingclaims thatthis is also a stylised description of a mechanism that works alongside other mechanisms in the real world, tending to induce non-differentiation there too. In other words, *he claims that* his model is true in the abstract. But that truth claim is just a reformulation of the claim that the model is explanatory. If one asks *what makes* Hotelling’s model an explanation of non-differentiation in the real world, it is no answer to be told that Hotelling thinks that his model represents a mechanism that is truly in operation.

The claim that a model is ‘true in the abstract’ might be understood in a different sense, exemplified by Nancy Cartwright’s (1998) account of how we can use models to learn about capacities. Hotelling’s analysis would fit Cartwright’s account if the proposition that the non-differentiation mechanism operates *in the real world* were derived from ‘accepted principles’ of economics by deductive logic, and if that derivation were accomplished by the proof of the results of the model. Then one *could* answer the question ‘What makes the model explanatory?’ by saying that the model is true in the abstract. But, as Reiss explains, Hotelling’s analysis cannot meet Cartwright’s conditions. Nor, as I have argued before, can most other economic models (Sugden 2000, 2009). The problem is that the implications that the modeller wants his readers to draw about the real world cannot be deduced merely from the accepted principles of economics: what *can* be so deduced are implications only about the model world. Cartwright (2009) recognises this, but sees it as revealing the unsatisfactory state of economic theory. In the sense that is relevant for a discussion of how models explain, economic models are false.

Some methodologists argue that simple and abstract economic models such as Hotelling’s are not intended to be explanatory. In one version of this argument, models are understood as potential explanations of regularities that would be observed in the real world, were the assumptions of the model to hold; the modeller does not make any claims about actual regularities in the real world (e.g. Aydinonat 2007; Grüne 2009). Other versions of the argument interpret some aspects of economic modelling as ‘conceptual exploration’ of the internal properties of theories (Hausman 1992, pp. 78–79), or as the investigation of recurring structures in economic interaction, without any necessary reference to specific cases (Schelling 2006, pp. 235–248; Ross 2008).

Certainly, if one relies on what economists say about their models, it is often difficult to be sure whether any particular model is intended to be explanatory or not, and if it is, what it is supposed to explain. Typically, an author says very little about how his model relates to the real world. There is a fully-specified model world in which agents behave according to principles determined by the modeller; properties *of that world* (the ‘results’ of the model) are discovered by deduction. In parallel, there is an informal discussion of regularities in the real world that in some way resemble the model. But surprisingly little is said about the implications of these similarities (Sugden 2000, 2009, 2011a). The author can be read as claiming that the model explains the real-world regularities; but alternative readings are often possible.

For example, consider Hotelling’s model. On my reading, Hotelling does claim, as Reiss says he does, that his model provides a (perhaps provisional and tentative) explanation of such supposed facts as that business districts in American cities are ‘too concentrated’ and that different brands of cider are ‘too homogenous’. Since the minimum-differentiation result breaks down if there are more than two firms in the model, that claim is not particularly convincing in general (though it does fit one case mentioned by Hotelling, that of political competition in America between the Democratic and Republican parties). But this is not the main point of Hotelling’s paper. Hotelling presents his analysis as a reconciliation of two apparently conflicting models of oligopoly – those of Cournot and Bertrand. In Cournot’s model, each firm chooses the quantity it will supply to the market; in equilibrium, the market price is greater than marginal cost. In Bertrand’s model, each firm chooses its price; in equilibrium, both prices are equal to marginal cost. In its initial version (in which firms’ locations are fixed), Hotelling’s model introduces what is effectively product differentiation, represented in spatial terms. Hotelling shows that when there is product differentiation, Cournot’s equilibrium result holds even if, as in Bertrand’s model, firms set prices. This exercise might be construed as an explanation of pricing in real-world markets in which there are only a few competing firms, but it might equally well be interpreted as an exploration of the internal logic of oligopoly theory, or as an analysis of oligopoly as a recurring economic structure.

But although a particular model might be presented without any specific explanatory target in mind, the enterprise of economic modelling as a whole is surely supposed to lead to explanations of real-world phenomena. Hotelling’s readers could be expected to be interested in his theoretical reconciliation of the Cournot and Bertrand models because those models were considered to be viable explanations of real-world oligopoly behaviour. So the problem is just shunted on: What makes (say) the Cournot model explanatory? Perhaps one might claim that Cournot did not propose his model as an explanation of any specific economic phenomena; it just provided a starting point for the development of more complicated or more industry-specific oligopoly models, and it is these that actually explain. But then what mysterious additional ingredient turns non-explanatory models into explanatory ones? As far as I can see, there is no such ingredient: simple, canonical models such as Cournot’s and Hotelling’s exhibit the methodological properties of economic modelling in general. If economics as a whole is explanatory, we should expect its canonical models to be explanatory too.

So I endorse Reiss’s first two propositions: (most) economic models are fictions, but (some of) these fictions are nevertheless explanatory. Reiss creates a paradox by claiming that an explanation of some phenomenon must provide a true description of the causal mechanism that is responsible for it. My account of models as credible worlds contravenes that principle. I shall now defend that account against Reiss’s objections.

In large part, those objections are based on what seem to me to be misunderstandings. Reiss is right that, in my account, a model is ‘a deliberate construction, by the modeller, of ... a parallel or counterfactual world that, to a greater or lesser extent, resembles aspects of our own world’, and that we learn about the real world by making inductive inferences from the model world. But he begins to go wrong when he attributes to me the claim that ‘we may infer a model result to hold true of a real-world phenomenon ... only to the extent that the parallel world depicted by the model is “credible”’, and then says that an assessment of the credible-worlds account has to ask ‘whether “credibility” can act as a stand-in for explanatoriness’ (p. 55). He goes on to summarise my position as: ‘In [Sugden’s] view, a model that describes a world that is “credible” is one that is explanatory’ (p. 57). Reiss seems to be reading my account as the view that the credibility of a model is sufficient to establish it as explanatory. That is a mistake.

Rather than considering what I have meant by ‘credibility’, Reiss inserts his own definition: ‘There is something that characterises good economic models in virtue of which they are acceptable by the economics community. Let us call that their credibility’ (p. 56). At this point, there is a slippage in Reiss’s argument. I *do* claim that there is something – or rather, there are some things – that characterise good economic models and allow them to be judged explanatory within the community of economists; and my concept of credibility is one of them. But that is not to assert that, whenever an economic model is judged acceptable by economists, there is thereby reason (for people in general) to treat it as explanatory. So, although Reiss is probably right to say that economists prefer models ‘that are mathematised, employ the tools of rational choice theory and solve problems using equilibrium concepts’, my account of models does not imply these preferences provide ‘good reason for considering models with these characteristics as explanatory’ (p. 56).

In my 2000 paper (Sugden, 2000, pp. 19–20) , I describe the credible-worlds mode of explanation by the following schema of inductive inference, in which *R* is ‘a regularity ... which may or may not occur in the real world’ and *F* is ‘a set of causal factors... which may or may not operate in the real world’:

E1. In the model world, *R* is caused by *F*.

E2. *F* operates in the real world.

E3. *R* occurs in the real world.

*Therefore, there is reason to believe:*

E4. In the real world, *R* is caused by *F*.

I also describe a parallel schema of *abduction*, which I present as a possible reading of some celebrated exercises of economic modelling:

A1. In the model world, *R* is caused by *F*.

A2. *R* occurs in the real world.

*Therefore, there is reason to believe:*

A3. *F* operates in the real world.

Notice that these schemas can come into play only after a theorist has constructed a model that induces some regularity *R* that replicates (at least with some degree of similarity) a corresponding regularity in the real world, generating that regularity through the workings of some set of causal factors *F*. Anyone who has tried to build economic models will know that achieving this kind of fit with the world is not easy; the closer the fit, and the more constraints one imposes on the properties of an acceptable model, the more difficult it is to achieve. This goes some way to explain why the ‘Therefore ...’ can carry some conviction. The underlying thought is that, if one finds a surprising and salient pattern in one’s observations, the surprise and the salience alone give some reason to expect that pattern to be projectible (Sugden 2011b). In this case, the surprise is that it has been possible to replicate real-world regularities in a simple model that satisfies the recognised constraints of modelling practice.

The clause ‘Therefore, there is reason to believe ...’ would perhaps have been better expressed as ‘Therefore, there is *some* reason to believe ...’. In my 2000 paper, I do not claim that whenever E1, E2 and E3 (or A1 and A2) hold for some *R* and *F*, one is entitled to infer E4 (or A3). Instead, I ask what factors might influence one’s degree of confidence in those inferences. I argue that one such factor is the robustness of the result ‘In the model world, *R* is caused by *F*’ to variations in apparently inessential features of the model world (pp. 21–23). But I give particular emphasis to judgements of *similarity*: ‘if we are to make inductive inferences from the world of a model to the real world, we must recognize some significant similarity between those two worlds’ (p. 23). The greater the similarity between the model world and the real world, the more confidence we can have in inferences from the former to the latter. This is where credibility comes in: ‘I want to suggest that we can have more confidence in [such inferences], the greater the extent to which we can understand the relevant model as a description of how the world *could be*’ (p. 24, italics in original). The capacity of a model to be understood in this way is what, in the credible-worlds account of models, ‘credibility’ means. Contrary to Reiss’s interpretation, credibility is not just whatever economists happen to like to see in their models. And credibility is neither necessary nor sufficient for a model to be judged explanatory; it is just one important dimension on which the degree of similarity between a model and the real world can be assessed.

The fundamental explanatory concept in my account of models is not credibility but similarity. As I have acknowledged in my more recent papers, this account has many features in common with Ronald Giere’s (1988) account of scientific explanation, of which I was not aware when I wrote my 2000 paper. In a later paper (Sugden 2011a, p. 733), I summarise my understanding of how models explain as follows: ‘The model is a self-contained construct, which can be interpreted as a description of an imaginary but credible world. The workings of the model generate patterns in the model that are similar to ones that can be observed in the real world. The model provides an explanation of the world by virtue of an inductive inference: roughly, from the similarity of effects we infer a similarity of causes. Various commentators have objected ... that mere similarity is insufficient to support an inductive inference. But perhaps, as Giere’s account of science implies, there is nothing more to scientific explanation than finding similarities between models and real-world phenomena.’ To ask for some further reason for holding that a model ‘genuinely’ explains some phenomenon (as Reiss seems to do on p. 56) is to ask one question too many.

Of course, it is inescapable that judgements of similarity are to some degree subjective. If my account of how models explain is accepted, explanation must inherit this subjectivity: one person (or one scientific community) may judge that a particular model is similar to some aspect of the real world, while another person (or another scientific community) may judge otherwise. To judge that a model is similar to the real world in the sense that it ‘*could be* real’ is to accept that 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 world’ (Sugden 2000, p. 25, italics in original). What I think I know (and think you should know) may not be the same as what you think you know (and think I should know). For example, consider some theoretical model in which the set of causal factors *F* includes sophisticated rational choice by individuals. The workings of these factors induce a regularity *R* in the model world. The same regularity is known to occur in the real world. How far does this modelling result give one confidence that the real-world regularity is caused by real-world rational choice, in the way that the model describes? Suppose that economist *A* believes that most individuals, most of the time, make their decisions much like the agents of rational choice theory, while economist *B* believes that human decisions are often determined by heuristics and biases of the kind studied by psychologists. Then the degree of similarity between the model and the real world is greater as judged by *A* than as judged by *B*; accordingly, *A* will be more impressed by the result than *B* will be.

So Reiss is right to say that judgements of credibility can be affected by such factors as the judge’s experiences, values, upbringing and education, and the theoretical preferences and history of the community of researchers to which she belongs. The implication for my account is that what one person finds explanatory, another person may not. For Reiss, it seems, this is a fatal objection: experiences, values, theoretical preferences and so on ‘have no essential relationship with explanatoriness’ (p. 56). Reiss seems to be expecting philosophy of science to supply rational criteria for determining what is and is not an explanation, such that, by following those criteria, good science can certify the objective validity of its explanations. In contrast, I would be inclined to say that explanatoriness is ultimately a psychological notion. If inductive inferences can be grounded on subjective perceptions of similarity and salience, as I have argued (Sugden 2011b), then we should expect the same to be true of some aspects of scientific reasoning.

That is not to say that scientific knowledge is just a social construction. We can ask of any particular community of researchers, with its given history and its evolving pattern of characteristic modes of enquiry, theoretical preferences and similarity judgements, how far it has been successful in discovering unexpected but predictable regularities in its domain of enquiry. To the extent that a research community can show such success, *that* is its claim to credence and authority – not its compliance with principles of reasoning laid down by philosophers of science, logicians or decision theorists. As I have said in a previous paper (Sugden 2009, p. 19): ‘I am not claiming to contribute to the enterprise in which philosophers of science explicate abstract principles by which claims to knowledge or belief can be rationally grounded. My approach is both more naturalistic and more pragmatic. My aim is to investigate the modes of reasoning that economic theorists use in their work, and to assess whether these are effective in helping them to understand real economic phenomena. Since we all find it necessary to use inductive inferences in our everyday lives, it should not be surprising to find that they are part of the practice of science too – however problematic they may be for professional logicians.’

**References**

Aydinonat, N.E. (2007), ‘Models, Conjectures and Exploration: an Analysis of Schelling’s Checkerboard Model of Residential Segregation’, *Journal of Economic Methodology*, 14, 429–454.

Cartwright, N. (1998), ‘Capacities’, in *The Handbook of Economic Methodology*, ed. J. Davis, W. Hands and U. Mäki, Cheltenham: Edward Elgar, pp. 45–48.

––––– (2009), ‘If no Capacities, then no Credible Worlds. But Can Models Reveal Capacities?’ *Erkenntnis*, 70, 45–58.

Giere, R. (1988), *Explaining Science*, Chicago: University of Chicago Press.

Grüne-Yanoff, T. (2009), ‘Learning from Minimal Economic Models. *Erkenntnis*, 70, 81–99.

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

Hotelling, H. (1929), ‘Stability in Competition’, *Economic Journal*, 39, 41–57.

Mäki, U. (1992), ‘On the Method of Isolation in Economics’, *Poznań Studies in the Philosophy of Science and the Humanities*, 26, 316–351.

––––– (2009), ‘MISSing the World: Models as Isolations and Credible Surrogate Systems’, *Erkenntnis*, 70, 29–43.

Reiss, J. (2012), ‘The Explanation Paradox’, *Journal of Economic Methodology*, 19, 43–62.

Ross, D. (2008), ‘Ontic Structural Realism and Economics’, *Philosophy of Science*, 75, 732–743.

Schelling, T. (2006), *Strategies of Commitment and Other Essays*, Cambridge, MA: Harvard University Press.

Sugden, R. (2000),‘Credible Worlds: the Status of Theoretical Models in Economics’, *Journal of Economic Methodology*, 7, 1–31.

––––– (2009), ‘Credible Worlds, Capacities and Mechanisms’, *Erkenntnis*, 70, 3–27.

––––– (2011a), ‘Explanations in Search of Observations’, *Biology and Philosophy*, 26, 717–736.

––––– (2011b), ‘Salience, Inductive Reasoning and the Emergence of Conventions’, *Journal of Economic Behavior and Organization*, 79, 35–47.