Journal of Economic Methodology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rjec20

# The explanation paradox redux

Julian Reiss [a]
[a] Department of Philosophy , Durham University , Durham , DH1 3HN , United Kingdom

Published online: 23 Sep 2013.

To cite this article: Julian Reiss (2013) The explanation paradox redux, Journal of Economic Methodology, 20:3, 280-292, DOI: 10.1080/1350178X.2013.828874

To link to this article: http://dx.doi.org/10.1080/1350178X.2013.828874

*Journal of Economic Methodology*, 2013
Vol. 20, No. 3, 280–292, http://dx.doi.org/10.1080/1350178X.2013.828874

Routledge
Taylor & Francis Group

# The explanation paradox redux

Julian Reiss*

*Department of Philosophy, Durham University, Durham DH1 3HN, United Kingdom*

(*Received 1 April 2013; accepted 1 May 2013*)

I respond to some challenges raised by my critics. In particular, I argue in favour of six claims. First, against Alexandrova and Northcott, I point out that to deny the explanatoriness of economic models by assuming an ontic (specifically, causal) conception of explanation is to beg the question. Second, against defences of causal realism (by Hausman, Mäki, Rol and Grüne-Yanoff) I point out that they have provided no criterion to distinguish those claims a model makes that can be interpreted realistically (the model's 'causal content' or claims it makes about causal powers or mechanisms) and those the realist can safely ignore. Third, I point out that Hausman's and Rol's claims about robustness plus the empirical observation that economic models are hardly ever robust to the right kinds of specification changes imply that these models do not explain (which is problematic as my original article argued). Fourth, I point out that Sugden's response still leaves an important question unanswered, viz. what makes economic models explanatory. Fifth, I sketch an alternative, instrumentalist account of explanation and argue that it would fit Sugden's bill. Sixth, I point out that under this account of explanation, economic models would come out as non-explanatory, which brings us back to Alexandrova and Northcott's account and its associated difficulties. The 'explanation paradox', therefore, remains unscathed.

**Keywords:** economic models; scientific explanation; instrumentalism; causality

First, I would like to thank the editors of the *Journal of Economic Methodology*, John Davis and Wade Hands, for giving me and the commentators a forum to discuss economic models and their explanatory status so extensively. Models are a major instrument in the economists' toolbox and explanation is one of the more important goals of any science, so the ideas discussed here are of great significance in the philosophy of economics. I also want to thank Anna Alexandrova and Robert Northcott, Till Grüne-Yanoff, Dan Hausman, Uskali Mäki, Menno Rol and Robert Sugden for taking the time to think so hard about the issues raised by the 'Explanation Paradox' (EP, Reiss 2012c) and writing a response. Menno Rol deserves special thanks because he initiated the symposium. Naturally, I disagree with much of what the commentators say, but I believe that we have made some real progress with respect to some questions. Certainly I have learned a lot.

The EP pursued three aims simultaneously. It had, first, an educational aim. I wanted to discuss existing approaches to models in economics and had to organize them somehow. I could have done it differently, of course, but the paradox appeared to provide a neat narrative within which many of the more specific points about the existing approaches would come up quite naturally. None of the commentators' remarks concern this educational aim, so I will ignore it here.

The second aim was systematic. When I started to think about models in economics methodically, I felt that they perform some sort of epistemic hat-trick. They seem to give us

*Email: julian.reiss@durham.ac.uk

some new knowledge about socio-economic states of affairs without involving new observations or experiments. In that sense, they are much like thought experiments. So do they really teach us something new about the world outside? Can they confirm scientific hypotheses in the absence of new data? I have been interested in that question for some time now (e.g. Reiss, 2002, 2008, 2012a). EP intended to add a new twist to this debate by replacing 'no new observations or experiments' by 'false models' and 'confirmation' by 'explanation'. An epistemic hat-trick though it remains. Can models explain even though they provide derivations of explananda that make essential use of premisses known to be false?

The only set of comments taking issue with the epistemic hat-trick is Alexandrova and Northcott's. In line with their previous work on models, these authors deny that there is a hat-trick going on. What appears to be a rabbit is just that: a mere appearance of a rabbit, not a 'real' rabbit. Alexandrova and Northcott (A&N) are correct in pointing out that this appearance is itself explanation-seeking. Why does it seem to us – and many economists – that models explain their phenomena when in fact, as A&N hold, they do not? Why do appearances deceive us?

A&N suggest a number of possible answers. The first builds on Salmon's distinction between 'epistemic' and 'ontic' conceptions of explanation and holds that models provide at best the former kind of explanations. Epistemic 'explanations', however, are not explanations, they are mere appearances of genuine explanations. A&N do not say this explicitly but they have to assume it on pain of not making their point. One example for an epistemic conception of explanation is the reduction of surprise. Models may well reduce our surprise at certain phenomena, but in so doing, they merely appear to explain the phenomena, they do not genuinely explain them.

EP left the choice of conception of explanation deliberately open, for a number of reasons. One is the broadly naturalistic reason that I prefer not to tell economists from the outside what conception they ought to adopt. So what if their explanations do not satisfy some philosopher's model of what a good explanation should look like. Does that tell us that economists fail to provide genuine explanations or rather that the account of explanation fails?

Another is that accounts can be explanatory in a variety of different ways, and the adequacy of this or that conception of explanation can only be determined within specific contexts that include epistemic and practical goals and purposes of an inquiry. Without specifying the context, I would find it hard to defend a given conception of explanation in a way that does not beg the question.

In particular, EP is not trying to defend a causal conception of explanation, pace what A&N say about this. The paper begins with causal explanation partly because the causal account is widely regarded as successful, and partly because adopting the causal account makes the paradox especially vivid. But nothing in the paper relies on that account and, indeed, I deliberately formulated the paradox in non-causal terms.

I therefore find A&N's rejection of 'epistemic' conceptions of explanation question begging. Without arguing in more detail that epistemic explanations are faulty in one way or another, I find it hard to accept that, as a matter of principle, epistemic explanations are mere pseudo explanations, that all 'genuine explanations' are of the 'ontic' kind. Perhaps epistemic explanations are all economists need. Indeed, EP gives a number of considerations in favour of an epistemic conception and Robert Sudgen also defends one in his comments.

I should mention in passing that whatever the economists' preferred conception of explanation, it is probably not of the 'reduction of surprise' sort. Economists often delight in providing explanations (genuine or pseudo) that add to the surprise of their audiences. The subtitles of books of the 'economics-made-fun' genre provide some evidence:

A Rogue Economist Explores the Hidden Side of Everything (Levitt & Dubner, 2005); The Unconventional Wisdom of Economics (Landsburg, 2007). Of course, these are intended for a general audience, but I think they contain popularizations of principles also at work in academic economics.[1] All truly great economics papers aim to provide novel, unexpected accounts of familiar facts. To give just two examples, Akerlof's lemons model (Akerlof, 1970) explains the large price differential between new cars and 'almost new' cars not in terms of a preference for novelty – which would be a familiar thing – but rather in terms of an informational asymmetry which, until then, no-one had thought of and which therefore must have caused surprise to most economists; Banerjee's 'simple model of herd behaviour' (Banerjee, 1992) explains herd behaviour not in terms of irrational group following – which would be a familiar thing – but in terms of rational responses to signals involving an informational externality which again must have been surprising to many at the time.[2] Whatever else they do, economics explanations reduce the familiar to the unfamiliar, not the other way around.

A&N's second error hypothesis is something like wishful thinking: 'Whenever we observe something consistent with the stylized claims of an idealized model it is correspondingly all too tempting to leap to the conclusion that that thing is explained by the model' (emphasis original). But the more relevant question is the reverse one: does any model whose results are consistent with a phenomenon of interest also explain the phenomenon? And the answer economists give to this question is clearly no. That a description of the phenomenon of interest can be derived from the model's assumptions is a necessary but not a sufficient condition for the model's explaining the phenomenon. The model needs a host of other characteristics to be regarded as explanatory. Sugden defends his view that its similarity to the target is one of these characteristics in his comments. I will say more about this below. For now, whatever these characteristics are, in EP I tried to get a grip on the connection between these characteristics and explanatoriness. Why are models that are similar to their targets (or mathematized, equilibrium-concept employing) considered explanatory?

One way to solve a mystery is to deny that it is a mystery in the first place. A&N deny the connection between characteristics of models and explanatoriness as they deny that models which have the right characteristics are necessarily or generally explanatory. But this raises another mystery: why do economists want models that have the desired characteristics? Are they perhaps individually rational but collectively crazy, just as one of Banerjee's herds?

To avoid that conclusion, A&N would have to give an alternative account of what models do if they do not provide explanations. In earlier work (e.g. Alexandrova, 2008; Alexandrova & Northcott, 2009), they indeed give such an alternative account. The account holds, essentially, that models play a heuristic role in the preparation of experiments: they provide causal hypotheses which are to be subjected to further empirical testing.

I do not deny that some models can sometimes play that role in economics. The problem is rather that if that is all models do, the mystery is not resolved. Why do economists build complex, mathematically sophisticated models rather than, say, resort to creativity and intuition, crystal balls, hypothesis-generating algorithms or consciousness-enhancing drugs? All of these sources of inspiration would be a lot easier to come by, and some of these would be more fun, than doing the hard work of constructing and solving a model. To warrant their existence, models must do more than to provide hypotheses. They must have some genuine epistemic benefit. EP asks but does not assume that this benefit is explanation. If not explanation, however, there must be something else, and least thus far A&N owe us an answer.

A&N's third error hypothesis is that giving these kinds of 'explanations' (note that these are merely of the epistemic variety and, therefore, not genuine explanations) is simply fun. Perhaps here are the beginnings of a genuine alternative account of modelling. Models tell us not about the socio-economic systems they superficially appear to represent but rather about their modellers. Economists enjoy building complex, mathematically sophisticated models for instance because the latter demonstrate their technical prowess and thus make for a selective advantage. Once selected, the real explanatory work can go on outside of theoretical journals, in experimental laboratories, departments of applied economics or policy circles. Mathematical models in this reading are much like antlers or long tails. Viewed from outside, they appear to make their bearers' lives hard – who would want to carry 8 kg of clumsy weight on his head? Who would want to learn the maths required to get into the *Journal of Economic Theory*? – but males worthy of reproduction and economists worthy of explanation will have to be selected somehow and antlers and mathematical knack help in the process (and we might as well enjoy parading them!).

This is a possibility. Obviously, it is a rather cynical account. I will not reveal here what I really think of it. Let me just reiterate that EP attempted a more conservative, less Freudian approach in which economists' claims to the effect that they intend their models to do genuine explanatory work were taken at face value and tried to account for. A&N's comments certainly make clear that alternative accounts are possible.

The third and far and away most important goal was strategic. I have pursued an instrumentalist agenda for some time now (e.g. in Reiss, 2012b), and EP intended to cast further doubt on the viability of the realist project but without mentioning the debate explicitly. The bulk of the comments are aimed at defending the realist cause, specifically at defending causal realism, and so in what follows I will take issue with arguments given on behalf of this stance.

Most members of my generation of philosophers of science were taught that comprehensive forms of scientific realism are almost certainly false. A look to the history of science teaches us that even our best theories are likely to be false, despite their predictive and technological successes. All previous generations of theories have been replaced by better ones. Would not it be preposterous to think that we have got it right today of all times? Scientific practice teaches us that fundamental theories require approximations and adjustments, simplifications and idealizations in order to be predictively and technologically successful; therefore, their successes speak at best in favour of localized applications of the theories, not of the fundamental principles themselves. Sociology of science teaches us that few scientific debates are settled on the basis of evidence alone but as non-evidential factors are arbitrary from the point of view of truth, current scientific 'state of the art' is hardly likely to be the last word.

Scientific realism has retreated in consequence. Contemporary realists defend forms of partial realism. The idea behind all forms of partial realism is that anti-realist arguments are successful at best with respect to some but not with respect to all scientific claims. Some claims (fundamental theoretical claims, say) are likely to be false and will therefore be replaced in the future; others (claims about causal relations or claims about structure), by contrast, stand a better chance of being true and may therefore stay. Partial realism is often motivated by demonstrations that some aspects of science withstand popular anti-realist arguments. Inference to the best theoretical explanation may well be faulty; but inference to the most likely cause does not suffer from the same deficits. Theories, including their hypotheses about fundamental ontology, may well be subject to perpetual change; but their mathematical structures often remain across theoretical change.

Partial realists thus pursue a divide-and-conquer strategy (Psillos, 1996). For the divide-and-conquer strategy to be successful, we need a criterion that distinguishes which aspects of a theoretical system the realist should put his/her money on from those he/she can safely ignore. Importantly, the former aspects will have to survive anti-realist arguments whereas the latter will not.

The reasoning in the case of models is not the same but analogous. We already know that some of the model's assumptions (and thus implications) are false. We cannot seriously be realists about claims such as 'Vendors set up business along a line segment of infinite thinness.' A realist about models has to argue that these assumptions are somehow ignorable or inessential in giving an explanation whereas other assumptions really matter.

Hausman makes it plain that he pursues this strategy:

> Consider the Phillips machine, which Reiss discusses (49). It misrepresents its target in many ways, but it accurately represents the fact that savings reduce current consumption. If only that proposition from the use of the model figures in an explanation, then the Phillips machine can be used in a true explanation. The fact that a model may contain many falsehoods does not preclude employing the model in giving explanations if the explanations do not rely on the falsehoods.

Mäki, similarly, writes that his account of models is a 'functional decomposition account .... It is a decompositional account since it relies on splitting models into bits and pieces rather than dealing with them as undifferentiated wholes'. EP argued that no such distinction can be made. There simply is not a criterion that enables us to parse the model's assumption into essential and inessential components.

Grüne-Yanoff, Hausman, Mäki and Rol all argue that the model's essential components are claims about isolated causal factors or 'mechanisms'. But which of the assumptions in models such as Hotelling's are about an isolated causal factor or 'mechanism'? Which assumptions 'isolate' the causal factor or 'mechanism' from disturbing factors?

Some of the comments discuss other examples than Hotelling's model or models like this. Grüne-Yanoff discusses Fehr and Schmidt's 1992 'model' of inequality aversion at some length (Fehr & Schmidt, 1999, p. 822):

$$U_i(x) = x_i - \alpha_i \max\{x_j - x_i, 0\} - \beta_i \max\{x_i - x_j, 0\}, \quad i \neq j,$$

where the $x$s are player $i$'s and $j$'s monetary payoffs and $\alpha$ and $\beta$ are parameters measuring how disadvantageous and advantageous inequality affects a player's utility.

In my own terminology (Reiss, 2013, 74ff.), Fehr and Schmidt's equation is not a model as such but rather a bridge principle that connects the concepts of theoretical models (in this case, those of game theory) with observable characteristics of situations of interest (which are in this case games played in laboratory environments). Game theory makes predictions about what will happen, given the utility of alternative courses of action. But utilities are not observable. So in order to apply game theory to a real situation such as a laboratory experiment, observable characteristics of the environment – such as monetary payoffs – have to be translated into utilities. That is what Fehr and Schmidt's equation does. It provides theoretical models such as the ultimatum and dictator games with empirical content so that these models can be used for predictions or explanations. On its own, it does not make any predictions whatsoever.

It does not matter of course whether one calls Fehr and Schmidt's equation a bridge principle (as I do) or a 'model for preference formation' (as Grüne-Yanoff might), what matters is that EP made claims only about a certain kind of model, namely theoretical models such as Hotelling's. That there are other models, in economics and in other sciences, that have different characteristics is of course true but beside the point. Yes, Galileo's model of the inclined plane isolates a causal tendency. What does this teach us about Hotelling?

Similarly with Rol's favourite example, the Madrid metro map. Maps can indeed misrepresent their targets in a myriad of ways as long as they retain certain topological properties. But these properties they should really have. If my Madrid metro map tells me I get to the airport from the Bilbao station by taking the brown line one stop South, change into the dark blue line at Tribunal, and then again at Nuevos Ministerios into the pink line, these relations and connections had better reappear in the actual metro system. By contrast, it does not matter that all lines are represented as having only right angles. But EP did not talk about maps either.

There is an important disanalogy between models and maps. The targets of maps are independently known, by surveys of the territory or simply because we built them ourselves, as the Madrid metro system. Map 'results' should therefore not surprise us. Their epistemic function is to guide the map user through the target which is unknown to him but known to the map maker. Scientific models, by contrast, chart new territory, territory that is otherwise inaccessible to the modeller. We build models (at any rate, those models EP talks about) because their targets are unknown and in order to learn something new about it.

So how do we know which of the claims contained in models like Hotelling's are to be taken seriously, checked for their truth by comparing them with socio-economic systems of interest and which we can ignore? Both Mäki and Rol say that only relevant features matter (Mäki; Rol). But they fail to provide a usable criterion of relevance.

In a map, and here is where the disanalogy really matters, because targets are known and the map purpose is well specified and simple, it is easy to determine relevance. Most maps are used for orientation, the Madrid metro map surely is, and so all those features are 'relevant' which help the map user to find out where he/she is and how to get from A to B. The label on the node three nodes down from Bilbao on the light blue line matters because I might want to know where I am after taking the metro South at Bilbao and getting off at the third stop. If the map is incorrect I will soon find out when I climb up the stairs and do not find myself at the Plaza del Sol.

EP talked about models whose targets are largely unknown and epistemically inaccessible to all of us, including the modeller, and scientific explanation was taken to be the modelling purpose. What, then, are the 'relevant' features of the model?

What we view as 'relevant' will certainly depend on our conception of explanation. Thus, under a deductive-nomological conception, everything necessary for the derivation of the explanandum will be relevant. This is an obvious non-starter for our problem because a typical model's many falsehoods could not be ignored (as they are necessary for the derivation of the explanandum – if they were not, why assume them?).

EP therefore does not discuss deductive-nomological explanation. EP takes causal explanation as the starting point because it is widely accepted as an adequate conception of explanation, as mentioned above, and because at least some accounts contain resources to help solve the present problem. Under a tendency, capacity or powers account of cause (potential differences between these views do not matter for the discussion here), causes sometimes continue to affect outcomes in systematic ways when disturbing factors are present. As is well known, forces in mechanics operate in this way. Hence, we often find mechanical examples in discussions of causation.

When causes operate in this way, certain idealizations are benign because, depending on the context, they make a negligible difference to the outcome or the idealized model makes a prediction about what would happen in the absence of the disturbing factor, which may either contribute to an explanation of an actual outcome or help in making a prediction about the actual outcome because the actual outcome is a result of a number of factors, the tendency of each of which and the law of composition are known.[3]

Idealizations that help to calculate what a causal factor does in isolation are called Galilean. EP argued at length that the typical idealizations typically found in economic models are not Galilean. I will not repeat the arguments given in EP here but I want to try to make the main point plain by again comparing the paradigmatic Galilean case – the free fall – with the non-Galilean case – Hotelling's law. Here are formulations of the two laws:

(1) The law of falling bodies states that falling body in a vacuum accelerates at the rate of 32 ft/s (9.8 m/s) during each second that it falls.
(2) Hotelling's law states that, in many markets, it is rational for producers to make their products as similar as possible.

On the face of it, the two laws are exactly parallel. Both predict what will happen under a *ceteris paribus* condition. The *ceteris paribus* condition in the first law reads 'in a vacuum', that in the second law, 'in many markets'. But the formally analogous *ceteris paribus* conditions are significantly different in content. The former says that when a certain disturbing factor (air resistance) is absent, the law will hold. I ignore the fact that it should also include other disturbing factors as the point will be exactly the same. The latter mentions otherwise unspecified 'many markets'. It does not say anything like 'in the cider, shoe and electronics market and in the markets for votes and devotion' or 'in markets where informational asymmetries don't matter'. So when we want to apply the law, when do we have reason to trust its prediction? The only way to find out is to consider Hotelling's model in which he derives the law. Unfortunately, the 'market' Hotelling describes is not one that can be found in the economy: one where exactly two producers are located on a line segment, consumers have perfectly inelastic demand schedules, transportation costs are linear and so on. Nothing in the model or the law tells us which of these assumptions are meant to be part of the *ceteris paribus* condition and which can be ignored. If they are all part of the condition, we know that the law is trivially true because not empirically instantiated. If none of them is part of the condition, we know that the law is false because there are markets in which rational producers do not make their products as similar as possible.[4]

Mäki is not happy with my distinction between 'assuming away' (e.g. air resistance) and 'assuming that' (e.g. there are exactly two producers located on a line segment) and argues that to 'assume that' usually implies to 'assume away' this and that. Thus, to 'assume that' competition is perfect implies for instance that the influence of price making and entry restrictions are 'assumed away'. This is a fair point I could have made clearer, and it is also rather trivial and does not change my main claim. Trivially, all models are partial representations of socio-economic systems. In that trivial sense, Hotelling's model 'isolates' some causal factors. Rationality seems to be an important factor and perhaps transportation costs. But Hotelling's model does not predict what rationality does in the presence of transportation costs in an otherwise vacuum. Rather, it tells us what rationality does in a situation where transportation costs assume a specific functional form, exactly two producers compete, set up business on a line segment, demand is inelastic and so on.

One might want to argue that the difference is one of degree, not of kind. The law of falling bodies calculates acceleration in terms of the gravitational constant and time – two factors, assuming away all the others. Hotelling's law predicts product differentiation behaviour in terms of, say, eight factors, assuming away all the others. But Galileo's assumptions only state that extraneous factors are not present; they do not ascribe strange properties or behaviours to the factors that are in the model. Hotelling's assumptions, by contrast, ascribe the factors in the model idealized properties and behaviours. It is as though the law of falling bodies assumed that the gravitational constant was 166 ft/s and that time was warped in addition to assuming that the bodies fall in a vacuum.

I am not saying that economics is alone in making assumptions of this type or that they are necessarily problematic. All I am saying is that a particular realist defence is inapplicable. The law of falling bodies can be held to describe what a falling body really does when other forces such as air resistance are absent. Hotelling's law cannot be held to describe what rationality and transportation costs really do when other factors are absent. It can at best be held to describe only what rationality and transportation costs were to do if they operated in highly special – and non-empirical – situation in which tons of factors are absent but in which in addition tons of factors operate in very specific ways. I have difficulty in understanding how one cannot see the difference.

Rol, I think, makes a related distinction but he does not show how it helps the realist cause. He writes that, 'What Galileo did is to apply a ceteris paribus clause; Hotelling, conversely, abstracted'. The difference seems to be that the former *explicitly* assumed away a causal factor (thus giving the *ceteris paribus* condition content) whereas the latter models a situation in ignorance of what factors might also affect the result (thus abstracting away from unknown factors).[5] It is not clear how pointing out that Hotelling ignores rather than assumes away extraneous factors helps with the present problem, which is, to wit, to find a criterion for parsing model assumptions into those whose truth value matters for the assessment of the model and those whose truth value is irrelevant.

EP also discusses robustness checks as a possible way out of the quandary. If model results were robust to specification changes, and the range of specification changes included approximately realistic conditions, then we could ignore the assumptions under which the result is robust and empirically test with respect to the remaining assumptions.

This point was taken up by Grüne-Yanoff. He writes that 'the question of stability [i.e., robustness] is ultimately an empirical one'. Of course I agree. So EP adduces a number of entirely empirical facts about robustness tests. First, they are hard to come by. Second, if feasible, they often show that the model result is not robust to specification changes. Third, in the rare cases in which model results are, or appear to be, robust to specification changes, this kind of robustness does not demonstrate that assumptions can be ignored.

Being generalizations, these facts can certainly be challenged. My 'demonstration' of facts one and two heavily relied on the literature that appeared after Hotelling's paper, and perhaps Hotelling's model is not very representative. I think it is, and if I had world enough and time, I would provide analogous results for a larger set of models. I also think that facts one and two do not require much further substantiation as they are well known. Everyone who has engaged in mathematical modelling in economics knows how hard it is to get any set of assumptions together from which a desirable result follows. And note that robustness tests are very demanding. Essentially, we have to calculate model results varying assumptions one by one and with respect to all possible alternatives. Does the Hotelling result hold when there are three, four, five or a continuum of producers? Assuming that there are three producers, does it hold when they are located on a rectangle, circle, globe or trefoil knot? Assuming that there are four produces who move on a rectangle, does it hold when transportation costs are quadratic, cubic or logarithmic? Assuming a continuum of producers moving on a Möbius strip with cubic transportation costs, does the result hold when demand elasticity is varied? The combinations are endless, and each one of these results will be hard if not impossible to calculate.

That model results tend to be sensitive to the specification is also a well-known fact. Most famously, it was described by Deirdre McCloskey as the A-Prime/C-Prime Theorem. Here is a statement of the Theorem (McCloskey, 1993, p. 235):

> For each and every set of assumptions A implying a conclusion C, there exists a set of alternative assumptions A′, arbitrarily close to A, such that A′ implies an alternative conclusion, C′, arbitrarily far from C.

Perhaps the results about Hotelling's model I gave in EP are unrepresentative, my experience is biased and so is McCloskey's. Of course these are possibilities. But the evidence speaks in our favour.[6]

What I meant by fact number three is that to the extent that model results are relatively insensitive to specification changes, the changes are often not significant or wide ranging enough or of the right kind to speak of genuinely robust model results. Just as there can be an A′ that is arbitrarily close to A but results in an alternative conclusion C′, there may well be an A″ that results in the same conclusion C. For a genuine robustness test, however, we would need a full permutation of assumptions (see above). So the existence of some A″ simply does not cut any ice. Moreover, it is also a fact that those A″ that have the robust conclusion C are often not arbitrarily close to the original A but rather significantly different.[7]

Rol and Hausman draw implications about the status of economics where robustness tests were unsuccessful. Rol writes:

> [Hotelling's reasoning] stands in a firm tradition of economists who investigate how a theorem, true under very strict conditions, fares if the conditions are weakened. If it holds, theorising is explanatory progressive. If not, then the research is degenerate.

In a similar vein, Hausman writes:

> Either there is robustness and the true claims concerning causes and mechanisms are doing the explaining, or there is no explanation.[8]

Since both Rol and Hausman accept premiss three of the paradox, namely that only true accounts explain, they reduce my paradox to a dilemma: either there is robustness and there is explanation, or there is neither. If what I say about robustness is true, both must be taken to deny that these models explain. This brings us back to A&N's response and its associated difficulties.

Hausman actually qualifies this claim. Whether or not a model's results are robust does not directly affect the model's explanatoriness. Failure to demonstrate that a model's results are robust might only make it hard to judge whether or not a model explains but whether or not it is indeed explanatory depends on whether 'it in fact identifies causes and relevant mechanisms'. Thus, as long as a model 'identifies a significant mechanism' and that mechanism operates in the world and is responsible for the result, it does not matter whether there are also falsehoods in the model.

However, also in this reading, the response collapses into a version of the first (the denial that economic models are explanatory). As mentioned above, target systems can, by and large, not be independently examined. This is why we resort to modelling in the first place. If we knew already what mechanisms are responsible for outcomes, we could of course build models that demonstrate their operation and use them for illustrative or educational purposes, perhaps like the Phillips machine. But we do not know the mechanisms independently, and models play an important epistemic role. One might deny this, but one would then have to explain why so much energy is expended on building models with surprising results (see above).

Robert Sugden defends a third response to the paradox, which says that economic models are explanatory, albeit in a way that does not require them to be true. While his comments make his position a lot clearer to me, I still think there is something missing from his account, namely a criterion enabling us to tell whether a given explanation is a good or adequate one or not. My original worry thus remains but it has moved to a slightly different place.

What EP failed to see is that Sugden regards models as explanatory not qua credibility but rather qua their similarity with target systems of interest: 'The fundamental explanatory concept in my account of models is not credibility but similarity'. Sugden then quotes a recent paper of his explaining his account (Sugden, 2011, p. 733), which is useful to repeat here in full length:

> 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 in 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.

Let me first say where I agree with this account: modelling involves an inductive inference – it is an ampliative mode of reasoning unlike the use of maps. But I continue to maintain that an important ingredient is missing from the account. First, the analogical inference Sugden describes faces an enormous underdetermination problem. In principle, the number of models in which the same phenomenon appears as result is unlimited. But they should not all be held to equally explain the phenomenon. To give a concrete example, consider Howard Davies' book on the recent financial crisis (Davies, 2010). Davies discusses 38 or so causes of the crisis, and when there are no constraints on what counts as an acceptable model and with some ingenuity, it would not be too difficult to build 38 different models, all of which 'explain' the crisis. Such an indiscriminate heap of models would not be a useful thing to have.

For one thing, not all models are compossible as explanations. Depending on the kind of regulation referred to, it may for instance not be possible that 'there wasn't enough regulation' and 'there was too much regulation' both explain the crisis. For another, some factors may be highly significant – from a theoretical, practical, regulatory and predictive … point of view – whereas others may at best be sidekicks. There must be a difference between explanation 4: US monetary policy and explanation 37: video games.

Not all these models are equally 'similar' to the socio-economic systems they represent. Some will make use of the tools of rational choice theory while others use those of behavioural economics (Sugden's example); I have seen neoclassical, Austrian, Marxist, Keynesian and other heterodox accounts of the crisis. Depending on the economist's judgements of what she thinks 'could be real' (*ibid.*), she will regard this or that model as more or most similar to the socio-economic system she seeks to explain.

Sugden makes plain that (a) the account he gives of explanation is subjective (or, in Salmon's terms, epistemic); and (b) subjective judgements of similarity are all there is to explanation. Who asks more of an account of explanation asks 'one question too many' (*ibid.*). I have no qualms with (a). As explained above, EP does not begin with a preconception on what should count as an adequate explanation, and especially it does not hold that good explanations must be of the 'ontic' type.

But while 'subjective' might not have to be bad, 'arbitrary' most certainly is. Even a naturalistic and pragmatist philosopher of science will want to provide tools that help resolve scientific disagreements. For him, the idea that 'one person … may judge that a particular model is similar to some aspect of the real world, while another person … may judge otherwise' (*ibid.*) will not do. EP tried to fill in this gap by asking whether unifying power may play the role of arbiter between competing accounts. Specifically, EP asked whether models that use principles that have higher unifying power should be regarded as more explanatory. According to Sugden, this was asking one question too many.

However, in fact Sudgen himself asks that question. He does not think that scientific knowledge is arbitrary. Rather,

> 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…. (emphasis original)

It is straightforward to turn these remarks into the beginnings of an instrumentalist theory of scientific explanation. Every epistemic community has a variety of goals, cognitive and practical. Explanation is a major cognitive goal. Sugden mentions the practical goal 'discovery of unexpected but predictable regularities'. Call that predictive success. There are further practical goals. Economics does not only want to uncover new phenomena, it also wants to help control known phenomena by successful interventions and policies. Call that strategic success. The instrumentalist theory of explanation holds that practical goals are primary and cognitive goals derivative. Accordingly, a scientific community is justified in regarding a model as explanatory to the extent that the model has exactly those characteristics which, in the long run, ensure (or contribute to) the epistemic community's predictive and strategic successes.

This is certainly a very attractive account of explanation. The problem is only that it does not resolve the paradox. Many commentators on economic practice, including many economists themselves, have taken the financial crisis as an opportunity to criticize contemporary economics for its failure to achieve its practical goals. Colander et al. say this in particularly plain English (Colander et al., 2009, p. 2):

> The global financial crisis has revealed the need to rethink fundamentally how financial systems are regulated. It has also made clear a systemic failure of the economics profession. Over the past three decades, economists have largely developed and come to rely on models that disregard key factors – including heterogeneity of decision rules, revisions of forecasting strategies, and changes in the social context – that drive outcomes in asset and other markets. It is obvious, even to the casual observer, that these models fail to account for the actual evolution of the real-world economy. […] There has also been little exploration of early indicators of system crisis and potential ways to prevent this malady from developing. […] Most models, by design, offer no immediate handle on how to think about or deal with [systemic crises]. In our hour of greatest need, societies around the world are left to grope in the dark without a theory. That, to us, is a systemic failure of the economics profession.

The models of contemporary economics would end up as not being explanatory because they are not useful to predict and prevent events such as the recent financial crisis. Thus, like the unification account EP discussed, this instrumentalist account of explanation does not help solve the paradox.

Here is a final suggestion. Perhaps this account of explanation has the resources for a more satisfactory error theory than those proposed by A&N. What matters for the account is success at prediction and policy in the long run. We will, of course, never know for sure what will happen in the long run. It takes events such as the recent financial crisis to draw our attention to the possibility that our current modelling practice is not quite as successful at our practical goals as we have thought or would have liked. It therefore draws our attention to the possibility that current models are not explanatory after all. Of course, the crisis does not prove anything. Perhaps we will just have to wait a bit longer until models have been constructed that have spectacular practical successes, even though they share the same characteristics as our current models. But it does cast a doubt on our intuition that these models explain.

## Notes

1. On the 'economics made fun' genre, see the symposium in *Journal of Economic Methodology* 19(3).
2. On 'herding' and the economics profession, see the symposium in *Journal of Economic Methodology* 20(1).
3. For an interpretation of idealizing assumptions explicitly along these lines, see Boumans (2003).
4. In an earlier work, Dan Hausman suggested that it might be hard to find a general characterization of the domain of application of a *ceteris paribus* law: 'I do not know whether there is a great deal to be said in an abstract way about what are appropriate specifications of domains of scientific investigation' (Hausman, 1992, p. 140). The point I am making here is that, even in a concrete case, it appears as though little can be said about the appropriate specification of a *ceteris paribus* clause.
5. This in fact strikes me as historically incorrect because Galileo (knew about but) did not explicitly mention air resistance in the Discorsi.
6. McCloskey's column from which the quote was taken ends as follows:

    > But the A-Prime/C-Prime Theorem proves that we [i.e., economists] are overinvesting in math-department questions of existence and underinvesting in physics-department questions of magnitude.
    >
    > Well, not exactly "proves". It's an empirical question. Come to think of it, I haven't yet found a proof of the Theorem. But as the physicist said, "You can whip up theorems; I leave that to the mathematicians."

7. Grüne-Yanoff misunderstands me when he says that 'I am sceptical about Reiss' claim of a strong link between non-Galilean assumptions and non-stability'. I nowhere make such a claim. The point about non-Galilean assumptions was that a certain realist defence is not available. If model results were robust, this would save the defence, but they are not. Other than that I did not claim any link between non-Galileanness of an assumption and the lack of robustness with respect to changes in the assumptions.
8. Hausman ascribes this view to me but this is a mistake. I did say, in the context of discussing robustness, 'Consequently, if these assumptions are false of an envisaged target system, we cannot expect the causal power or mechanism to operate in the system. This is detrimental to our explanatory endeavour' (Reiss, 2012c). However, in that context, I was considering only causal explanation because that is the conception proponents of the 'Models are true after all' solution to the paradox defend. I did not mean to say that models cannot be explanatory in a different, non-causal way.

## References

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

Alexandrova, A. (2008). Making models count. *Philosophy of Science*, *75*, 383–404.

Alexandrova, A., & Northcott, R. (2009). Progress in economics: Lessons from the spectrum auctions. In H. Kincaid & D. Ross (Eds.), *The Oxford handbook of philosophy of economics* (pp. 306–336). Oxford: Oxford University Press.

Banerjee, A. V. (1992). A simple model of herd behavior. *Quarterly Journal of Economics*, *107*, 797–817.

Boumans, M. (2003). How to design Galilean fall experiments in economics. *Philosophy of Science*, *70*, 308–329.

Colander, D., Föllmer, H., Haas, A., Goldberg, M., Juselius, K., Kirman, A., … Sloth, B. (2009). *The financial crisis and the systemic failure of academic economics*. Discussion Paper 09-03. Department of Economics, University of Copenhagen.

Davies, H. (2010). *The financial crisis: Who is to blame?* Cambridge: Polity Press.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*, 817–868.

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

Landsburg, S. (2007). *More sex is safer sex: The unconventional wisdom of economics*. New York, NY: Free Press.

Levitt, S., & Dubner, S. (2005). *Freakonomics: A rogue economist explores the hidden side of everything*. New York, NY: William Morrow.

McCloskey, D. (1993). Other things equal. *Eastern Economic Journal*, *19*, 235–238.

Psillos, S. (1996). Scientific realism and the ‘pessimistic induction’. *Philosophy of Science*, *63* (Suppl), S306–S314.

Reiss, J. (2002). *Epistemic virtues and concept formation in economics*. (Unpublished PhD dissertation). London School of Economics.

Reiss, J. (2008). *Error in economics: Towards a more evidence-based methodology*. Abingdon: Routledge.

Reiss, J. (2012a). Genealogical thought experiments in economics. In J. R. Brown, M. Frappier, & L. Meynell (Eds.), *Thought experiments in science, philosophy, and the arts* (pp. 177–190). New York, NY: Routledge.

Reiss, J. (2012b). Idealisation and the aims of economics: Three cheers for instrumentalism. *Economics and Philosophy*, *28*, 363–383.

Reiss, J. (2012c). The explanation paradox. *Journal of Economic Methodology*, *19*, 43–62.

Reiss, J. (2013). *Philosophy of economics: A contemporary introduction*. New York, NY: Routledge.

Sugden, R. (2011). Explanations in search of observations. *Biology and Philosophy*, *26*, 717–736.