Journal of Economic Methodology,
Vol. 19, No. 1, March 2012, 43–62

Routledge
Taylor & Francis Group

The explanation paradox

Julian Reiss*

Department of Philosophy, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

This paper examines mathematical models in economics and observes that three mutually inconsistent hypotheses concerning models and explanation are widely held: (1) economic models are false; (2) economic models are nevertheless explanatory; and (3) only true accounts explain. Commentators have typically resolved the paradox by rejecting either one of these hypotheses. I will argue that none of the proposed resolutions work and conclude that therefore the paradox is genuine and likely to stay.

Keywords: models; explanation; idealisation

1. Introduction

The causal account of explanation is widely regarded as successful and, importantly, more successful than its alternatives – both as an account of scientific explanation in general and one of explanation in economics in particular. To explain a specific economic event is to cite its causes; to explain a general economic phenomenon is to describe the causal mechanism responsible for it.[1]

The starting point for this paper is the observation of a particular feature of causal explanations: causal explanations cannot be successful unless they are true. I took this idea from Nancy Cartwright who, albeit to make a different point, wrote:

> My newly planted lemon tree is sick, the leaves yellow and dropping off. I finally explain this by saying that water has accumulated in the base of the planter: the water is the cause of the disease. I drill a hole in the base of the oak barrel where the lemon tree lives, and foul water flows out. That was the cause. Before I had drilled the hole, I could still give the explanation and to give that explanation was to present the supposed cause, the water. There must *be* such water for the explanation to be correct. An explanation of an effect by a cause has an existential component, not just an optional extra ingredient. (Cartwright 1983, p. 91; emphasis in original)

Cheap money in the early 2000s does not explain the financial crisis of the late 2000s unless money was indeed cheap (in the sense that interest rates were lower than the rate that would have been 'adequate' given the economic conditions), and unless cheap money was indeed the factor without which the financial crisis would not have occurred. The monetary transmission mechanism (or a description thereof) does not explain the aggregate relationship between money, the interest rate, and real variables unless changes in real variables are, at least sometimes, brought about by the transmission mechanism.

The requirement that causal accounts be true to be explanatory is in fact the great downside of causal explanation. When phenomena are complex, and economic

*Email: reiss@fwb.eur.nl

ISSN 1350-178X print/ISSN 1469-9427 online

http://dx.doi.org/10.1080/1350178X.2012.661069
http://www.tandfonline.com

phenomena are, truth is hard to come by. Accounts given of economic phenomena are usually dramatically simplified and features we know affect a result are represented in a systematically distorted way. Among economists, the slogan 'all models are wrong, but some are useful' (due to statisticians Box and Draper 1987, p. 424) is well-known. And yet, such models are regarded by economists and others as having more than heuristic value: not always, to be sure, but often enough economic models succeed in explaining.

The issue I aim to tackle in this paper is the question whether we can square the fact that all models contain significant falsehoods with the economists' aim to give genuinely explanatory accounts of economic phenomena: Do false models explain? To proceed, I will first describe, in Section 2, what I take to be a paradigmatic example for a false explanatory economic model in quite some detail. The amount of detail given will seem somewhat tedious initially but prove useful for the methodological points I make later on. In Section 3, I will then describe the various ways in which that model, and most other models in economics, is false. The main discussion occurs in Section 4 where I will formulate the problem as a paradox and classify the various responses given in the literature as denying a specific premiss in the paradox. Section 5 concludes quite soberly that since all the proposed resolutions fail, the paradox is genuine.

2. A model

Let us begin by briefly examining a classic example in the use of models in economics: Harold Hotelling's derivation of the principle of minimal differentiation which has become to be known as 'Hotelling's Law' (Hotelling 1929). Hotelling's starting point is the observation that if one of the sellers of a good increases his price ever so slightly, he will not immediately lose all his business to competitors – against the predictions of earlier models by Cournot, Amoroso and Edgeworth:

> Many customers will still prefer to trade with him because they live nearer to his store than to the others, or because they have less freight to pay from his warehouse to their own, or because his mode of doing business is more to their liking, or because he sells other articles which they desire, or because he is a relative or fellow Elk or Baptist, or on account of some difference in service or quality, or for a combination of reasons. (Hotelling 1929, p. 44)

The reason for this is that another economics law, the law of one price, is itself at best a *ceteris paribus* law. The law says that in one market the same goods must sell at the same price – if they did not, customers would flock to the cheapest seller, forcing more expensive sellers to lower their prices or driving them out of the market. But that of course holds only if the goods are identical in every respect, including their spatial distance to the buyer, which is never strictly true of actual goods. Hotelling's model describes what happens when one of the conditions in the *ceteris paribus* clause is relaxed: specifically, when goods differ in their spatial distance to the buyer along a single dimension.

Suppose, then, that the buyers of a commodity are uniformly distributed along a line segment of length l. Two vendors A and B are at distances a and b, respectively, from each end of the line segment (Figure 1).

The cost of production for the good to A and B is assumed to be zero. Demand is perfectly inelastic. Each buyer transports his purchase to the place where he consumes it at cost c per unit distance. Denote A's price by p_1, B's price by p_2 and let q_1 and q_2 denote the respective quantities.

Under these assumptions, B's price can exceed that of A without B losing *all* his customers to A. However, he must not let his price exceed that of A's by more than the transportation cost from A to B, which can be expressed as $c(l - a - b)$.

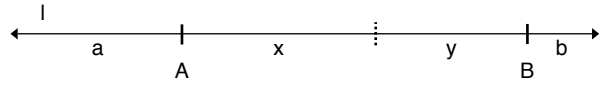

Figure 1. Hotelling's model of spatial aggregation.

In this way, he will attract all the business of the line segment to his right, b, plus some of the business in between A and B, which is denoted by y. The same is true for A of course, *mutatis mutandis*, so that A attracts all the business to his right plus some of the business in between A and B, denoted by x. The lower A's price as compared to B's, the more business can he attract, i.e., the greater is x.

The marginal customer is indifferent between A and B:

$$p_1 + cx = p_2 + cy.$$

Moreover, we know that:

$$l = a + x + y + b.$$

Solving for x and y, calculating profits = revenue = pq and substituting $a + x$ for q_1 and $b + y$ for q_2 yields:

$$\pi_1 = p_1 q_1 = p_1(a + x) = 1/2(l + a - b)p_1 - p_1^2/2c + (p_1 p_2/2c)$$

$$\pi_2 = p_2 q_2 = p_2(b + y) = 1/2(l - a + b)p_2 - p_2^2/2c + (p_1 p_2/2c).$$

Setting the derivative with respect to price to zero and solving gives the equations:

$$p_1 = c(l + (a - b)/3),$$

$$p_2 = c(l - (a - b)/3);$$

and:

$$q_1 = 1/2(l + (a - b)/3),$$

$$q_2 = 1/2(l - (a - b)/3).$$

Profits then are given by:

$$\pi_1 = p_1 q_1 = c/2(l + (a - b)/3)^2,$$

$$\pi_2 = p_2 q_2 = c/2(l + (b - a)/3)^2.$$

So far we have assumed that A and B have fixed locations. Let us now relax that assumption. It can readily be seen from the profit equations that A will want to make a as large as possible and B will want to make b as large as possible. That is, they will move towards each other. If, in the above figure, B moves first, he will immediately locate to the right of A. In this case, A will move to B's immediate right because that part of the line segment (in the figure; $x + y + b$) is larger than his segment on the left (a). Then B will move again to A's right and so on until they are both at the centre of the line, sharing the business equally.

It is important to assume that *A* and *B* cannot occupy the same position on the line because in this case they would enter into a price war, reducing profits for both. Hotelling remarks about this (Hotelling 1929, p. 52):

> From *B*'s standpoint the sharper competition with *A* due to proximity is offset by the greater body of buyers with whom he has an advantage. But the danger that the system will be overturned by the elimination of one competitor is increased. The intermediate segment of the market ($x + y > 0$) acts as a cushion as well as a bone of contention; when it disappears we have Cournot's case, and Bertrand's objection applies.[2]

The two will therefore move as close as possible to each other without becoming identical. This, argues Hotelling, is precisely what we observe in a large number of economic and non-economic phenomena (Hotelling 1929, pp. 54 and 57):

> In politics it is strikingly exemplified. The competition for votes between the Republican and Democratic parties does not lead to a clear drawing of issues, and adoption of two strongly contrasted positions between which the voter may choose. Instead, each party strives to make its platform as much like the other's as possible. [. . .]
>
> It leads some factories to make cheap shoes for the poor and others to make expensive shoes for the rich, but all the shoes are too much alike. Our cities become uneconomically large and the business districts within them are too concentrated. Methodist and Presbyterian churches are too much alike; cider is too homogeneous.

The 'too much alike' refers to the fact that the profit-maximising equilibrium differs from the social optimum in the model. Indeed, if *A* is located at one quarter of the line segment from the left and *B* at one quarter from the right, they would also divide the cake into half but customers would have to travel much less. But if *A* really did locate there, *B* would move to his immediate right, taking half of *A*'s profits.

3. Idealisations

It is obvious that Hotelling's model is highly idealised relative to the phenomena it seeks to explain. The most literal application of the model would probably be related to the location decisions of two businesses along a straight line such as shops on a Main Street or ice cream vendors along a beach. Even for such applications – and the model, as we have seen, is meant to apply much more broadly – the model makes numerous assumptions we know to be false: we move in three- and not in one-dimensional space; goods differ with respect to many aspects other than 'distance from point of consumption'; customers are not uniformly distributed along a line and demand is seldom completely inelastic; sellers act on numerous motives of which profit maximisation is at best one.

There are many classifications of different kinds of idealisations one might find in science. I find William Wimsatt's to be particularly useful (2007, pp. 101–102; emphasis in original):[3]

(1) A model may only be of very *local applicability*. This is a way of being false only if it is more broadly applied.
(2) A model may be an *idealisation* whose conditions of applicability are never found in nature, (e.g., point masses, the uses of continuous variables for population sizes, etc.), but which has a range of cases to which it may be more or less accurately applied as an approximation.
(3) A model may be *incomplete* – leaving out one or more causally relevant variables. (Here it is assumed that the included variables are causally relevant, and are so in at least roughly the manner described.)

(4) The incompleteness of the model may lead to a *misdescription of the interactions* of the variables which are included, producing apparent interactions where there are none ('spurious' correlations), or apparent independence where there are interactions – as in the spurious 'context independence' produced by biases in reductionist research strategies. Taylor (1985) analyses the first kind of case for mathematical models in ecology, but most of his conclusions are generalisable to other contexts. (In these cases, it is assumed that the variables identified in the models are at least approximately correctly described.)
(5) A model may give a *totally wrong-headed* picture of nature. Not only are the interactions wrong, but also a significant number of the entities and/or their properties do not exist.

On pain of trivialising (1), we should probably qualify 'in its intended domain'. No model explains everything; a model is always a partial representation of the world. But it is a substantial point to say that models often have local applicability, even in their intended domain. Unfortunately, it is neither quite clear what a model's 'intended domain' is nor what 'applicability' means.

Hotelling's paper gives some indication about where he intends his model to apply. He wants to draw our attention to the fact that consumers often deal with one seller rather than another one despite a difference in price. He explains that by product differentiation. This suggests that the intended domain is economic settings in which producers can erect quasi-monopolies by differentiating their product from competitors and can set prices in the light of maximising profits. This cannot be the end of the story because party politics is clearly within Hotelling's intended domain but parties at best maximise votes rather than profits, but let us ignore that here. What might it then mean for a model to be applied? Supposedly, it means to use the model to explain phenomena of interest and make predictions. The model might be falsified in this particular way if, for instance, two businesses do not compete via prices even though they could, or if they ended up in a price war because they produced identical goods.

That the model idealises in sense (2) is clear, among other things, from the fact that its two producers move along a line that has no breadth or thickness. How significant such an idealisation is depends on purpose and context. Hotelling himself sees a zero-dimensional market in economics in analogy to point masses in astronomy (Hotelling 1929, p. 45):

> To take another physical analogy, the earth is often in astronomical calculations considered as a point, and with substantially accurate results. But the precession of the equinoxes becomes explicable only when account is taken of the ellipsoidal bulge of the earth. So in the theory of value a market is usually considered as a point in which only one price can obtain; but for some purposes it is better to consider a market as an extended region.

Moving from a zero-dimensional geography in which the law of one price holds to a one-dimensional geography where Hotelling's principle holds is the minimum adjustment he could make. Whether considering a market as a line is a harmless idealisation depends on what aspects of the geometry are relevant for consumer decisions. It is often useful to consider cities as two dimensional. The shortest distance from A to B in two-dimensional space is of course a straight line or, when one cannot move in a straight line because of buildings and traffic, the closest approximation to a straight line. But when one travels by bike and the city is very hilly, such as La Paz or San Francisco, one usually fares better by taking the contours into account. Similarly, ice cream vendors might have to take account of the breadth and gradient of the beach if these geographical features matter to consumers.

Hotelling's model is also false in sense (3). Customers care about much more than how far they have to travel to get a product, no matter what the geography. Hotelling mentions various examples himself: the sweetness of cider, whether the seller is a fellow Elk or Baptist, party ideologies. When producers can differentiate their goods with respect to more than one characteristic, whether all the different characteristics can usefully be captured in a single transportation cost parameter depends on whether the different characteristics interact in their bringing about the outcome. Do ice cream vendors still move as closely together as possible when they can both change their location as well as the taste of the ice cream they sell?

Under (4) I would include assumptions to the effect that causal relations have specific functional forms in the absence of evidence that the modelled phenomena satisfy these functional forms. Transportation costs are assumed to be linear; consumer demand is assumed to be perfectly inelastic. These are at best approximations but more probably significantly wrong. Hotelling considers the case of elastic demand (Hotelling 1929, p. 56): 'With elastic demand the observations we have made on the solution will still for the most part be qualitatively true; but the tendency for B to establish his business excessively close to A will be less marked'. He asserts this without providing much evidence, however, and a result in which 'B will definitely apart from extraneous circumstances choose a location at some distance from A' (Hotelling 1929) is arguably a qualitatively different result than the principle of minimum differentiation.

A model is false in sense (5) when it gives a totally wrong-headed picture of nature, when the posited entities or properties do not exist. In economics this is a tricky type of idealisation as the entities and properties it posits always have counterparts in our everyday ontology of the world. Economics does not explain phenomena by introducing strange things such as electrons, quarks and strings, the id and the unconscious, *l'Élan vital* and *la Volonté generale*. Rather, ordinary things such as households and firms, businessmen, their plants and the goods they produce are transformed into something no less strange but with a clear analogue in everyday life. Typical economics models, let us say, assume businessmen to have perfect calculation abilities and care only about profits. But they are still businessmen. Thus, in some sense, even if all actual businessmen were particularly bad at maths and cared mostly about world peace, these models would not give a totally wrong-headed picture of nature.

I would nevertheless say that an idealisation falls into this category whenever the outcome of interest – say, minimal product differentiation – is produced by a causal mechanism that differs from the mechanism represented in the model. In the case at hand, the mechanism includes a conscious product differentiation on the part of businesses aiming to create a spatial monopoly to maximise profits. Minimal product differentiation could be a result of other mechanisms – imitation, say, or chance – which may or may not be aimed at profit maximisation. To the extent that such other mechanisms are at work, Hotelling's model gives a 'totally wrongheaded picture of nature'. To the extent, for instance, that politicians actually believe in the rightness of their politics, minimal differences between parties (where they exist) are misrepresented by Hotelling's model as being the result of a process of maximisation.

4. Explanation

Hotelling's model, then, is false in all relevant senses from (1) to (5) from Wimsatt's list. And yet, it is considered explanatory. Moreover, and perhaps more importantly, it feels explanatory. If we have not thought much about Hotelling's kind of cases, it seems that

we have genuinely learned something. We begin to see Hotelling situations all over the place. Why do electronics shops in London concentrate in Tottenham Court Road and music shops in Denmark Street? Why do art galleries in Paris cluster around Rue de Seine? Why have so many hi-fi-related retailers set up business in Calle Barquillo in Madrid such that it has come to be known as 'Calle del Sonido' (Street of Sound)? And why the heck are most political parties practically indistinguishable? But we do not only come to see that, we also intuitively feel that Hotelling's model must capture something that is right.

We have now reached an impasse of the kind philosophers call a paradox: a set of statements, all of which seem individually acceptable or even unquestionable but which, when taken together, are jointly contradictory. These are the statements:

(1) Economic models are false.
(2) Economic models are nevertheless explanatory.
(3) Only true accounts can explain.

When facing a paradox, one may respond by either giving up one or more of the jointly contradictory statements or else challenge our logic. I have not found anyone writing on economic models who has explicitly challenged logic (though their writings sometimes suggest otherwise). There are authors, however, who resolve the paradox by giving up a premiss. I will discuss one or more examples for each.

4.1 Economic models are true after all: in the abstract

Let me begin with a disclaimer. I do not think that models have true values. Whatever models are, and there is some debate about the 'ontology of models' (see for instance Frigg and Hartmann 2006), it is most certainly not the case that models are sentences. But its sentences that are true or false. For a very intuitive example, take the Phillips machine (or Monetary National Income Analogue Computer, MONIAC). The Phillips machine is a model of the UK economy. It consists of a number of interconnected transparent plastic tanks and pipes, each of which represents an aspect of the UK economy. The flow of money through the economy is represented by coloured water. The Phillips machine is not true or false, in the same manner as a tree is not true or false. Statements are true or false *of* the Phillips machine – for example that its tanks and pipes are mounted on a wooden board – just as statements are true or false of a particular tree. And this remains the case whether the model is an analogue or physical model, or whether it is a mathematical or otherwise abstract model.

Consequently, when we say that a model is true or false, we speak elliptically. Suppose that when Bill Phillips built his machine, it was a representation of the UK economy that was adequate for his purposes. For instance, in the Phillips machine one can reduce expenditure by draining water from the pipe that is labelled 'expenditure' and diverting it into a pipe that says 'savings'. This is an accurate representation in so far as savings reduce the funds available for expenditures in the UK economy. When I quoted the slogan 'all models are false' approvingly above, I meant to draw attention to the undisputed fact that all models also *mis*represent their targets in a myriad of respects. Whatever money in the UK economy is, it is not wet as the water in the Phillips machine. Whatever banks are, they are not plastic tanks filled with water. And so on. Thus, when we say colloquially 'all models are false' what we mean is 'all models misrepresent their targets in one way or another'. In case of an abstract model, we may alternatively say that some of the assumptions that define the model, and therefore are necessarily true of the model, are false

of the target system of interest. As another alternative, we may say that a theoretical hypothesis, which states that some target system is like some model, is true or false.

Our first strategy to resolve the paradox is to claim that a model can be true despite, or even in virtue of, containing many falsehoods. More accurately, a model can misrepresent its target in some (presumably, inessential) respects to correctly ('truthfully') represent other (presumably, essential) respects. Another way of putting the issue is that models are true in the abstract: they do not represent what is true but rather what would be true in the absence of interferences. Nancy Cartwright (1989) developed such a view as a general perspective on science in great detail. The main advocate of a related view of models in economics is Uskali Mäki (e.g., 1992, 1994, 2005, 2009, 2011).[4]

The core idea is that models can be thought of as Galilean thought experiments (Cartwright 1999). In a Galilean thought experiment, an experimental situation is contemplated after mentally removing 'disturbing factors' – factors different from the main cause under investigation, which nevertheless affect the outcome. To discover an instance of the law of falling bodies, say, the thought experimenter imagines a situation that is free from all factors that affect a body's rate of fall except the gravity of the Earth.

Mäki calls this mental process 'isolation by idealisation': the operation of one specific causal factor is isolated – in Galileo's case, the Earth's gravitational pull – by idealising away every other factor – air resistance, other gravitational fields, other forces. The resulting model is 'false' in many ways because these factors do affect all actual systems that we may choose as target systems of interest. But it is also 'true' in one important way: it correctly captures the operation of the causal factor of interest, the gravity of the Earth. Mäki makes this point with respect to von Thünen's model of the isolated state (Mäki 2011, p. 60):

> If there is a natural truth bearer here, it is neither this model as a whole nor just any arbitrary parts of it. It is rather a special component of the model, namely the causal power or mechanism that drives this simple model world: the Thünen mechanism. This truth bearer has a fair chance of being made true by its truth maker, the respective prominent causal 'force' or mechanism in the real system. It is the mechanism that contributes to the transformation of distance into land use patterns through transportation costs and land values.

In Mäki's parlance, then, models are not true *per se* but rather they may contain truths such as truths about the causal powers of mechanisms of the target systems of interest. It would probably be more accurate to say (for instance) that a theoretical hypothesis stating that the model correctly represents a target system's causal power or mechanism can be true, but let us not get drawn away by a trifle.

This line of defence is perfectly legitimate for a variety of false modelling assumptions in science. In many domains of science, especially in mechanics, has the method of analysis and synthesis been used with great success. Natural systems often do not obey to neat scientific laws because they are too complex and too changing. So we experimentally create – in the lab or in the head – situations that are simpler, more manageable and free from outside influences. We learn what happens in these situations and use that knowledge to predict what happens in more complex, more natural situations. This is often possible because what we learn in the simplified system remains true, with qualifications of course, in the more complex system. It is not an accident that Galileo is often regarded as, on the one hand, the originator of the idea that natural systems can be analysed as being composed of 'phenomena' – universal and stable features, which are of scientific interest – and 'accidents' – disturbing factors, which are not – and, on the other hand, the inventor of the world's most famous and most successful thought experiments (McAllister 2004).

At a first glance, economists are well aware of the method of analysis and synthesis, and regard their work as applications of this method. Our own Hotelling is a case in point.

As he says about the principle of minimum differentiation (Hotelling 1929, p. 54; emphasis added):

> But there is an incentive to make the new product very much like the old, applying some slight change which will seem an improvement to as many buyers as possible without ever going far in this direction. The tremendous standardisation of our furniture, our houses, our clothing, our automobiles and our education are due in part to the economies of large-scale production, in part to fashion and imitation. But *over and above these forces* is the *effect* we have been discussing, the *tendency* to make only slight deviations in order to have for the new commodity as many buyers of the old as possible, to get, so to speak, between one's competitors and a mass of customers.

Hotelling believes that his model does not represent a local causal principle with limited applicability – applicability only where the model's assumptions are met. Rather, it represents a more general tendency that persists (continues to affect outcomes) even in the presence of disturbing factors, which, in this case, are economies of scale, fashion and imitation.

The problem is only that the models of economics, Hotelling's included, are by and large very much unlike a Galilean thought experiment. Let us say with Mäki that a Galilean thought experiment isolates (the primary causal factor) by idealising (away other causal factors). Is this really what typical economics models do?

Few of the assumptions in Hotelling's model aim to eliminate disturbing causal factors. Assuming businesses move along a line with no breadth or thickness is not assuming away the influence of geography; it is determining a specific geography in which Hotelling's results are true. Assuming that transportation costs are linear in distances is not assuming away the influence of transportation costs; it is determining a specific functional form of the effect of transportation costs on utility. Assuming that demand is perfectly inelastic is not assuming away the influence of demand; it is determining a specific functional form of the demand schedule. And so on.

One might object that the distinction I am making here is spurious because to 'assume away' a causal factor is in fact a special case of the more general kind of idealisation just described. To 'assume away' air resistance is, so the objection goes, to assign a specific value to air resistance in the model – zero. Likewise, to assume that, say, transportation costs are linear is to assign a specific value to the transportation cost parameter in the model. However, there are at least three differences between Galilean and non-Galilean assumptions. First, in a Galilean thought experiment, the factor that has been 'assumed away' does not normally appear. The assumption of no air resistance cannot be read off the model. It only surfaces when we ask 'under what conditions would the result (given by the Galilean thought experiment) be true?' By contrast, the non-Galilean assumptions Hotelling uses are all an explicit part of the model, and they are assumptions without which no result could be calculated at all. That is, the assumption already appears when one calculates the model result, and not only when one uses the result to make a prediction about a phenomenon outside the model. Second, Galilean assumptions usually concern quantitative causal factors. Different media produce different degrees of resistance. Hotelling's assumptions are categorical. Different geographies are different kinds of thing and not the same kind of thing to a different degree. Third, Galilean assumptions usually concern a causal factor that has a natural zero. No air resistance is such a natural zero. Assuming that celestial bodies are point masses is another example: a point is the natural zero for the quantity 'extension'. Geographies and the functional form of transportation costs have no natural zero. The elasticity of demand may be considered to have a natural

zero ('perfectly inelastic demand') but that particular value still appears in the model, and therefore elasticity is not 'assumed away' but is rather part of the model.

The importance of making assumptions of the Galilean kind is made plain by the goal of a Galilean thought experiment, which is to learn what a causal factor does in isolation from disturbing factors (McMullin 1985; Cartwright 1989; for an application to social science, see Reiss 2008b). To 'assume away' air resistance in a thought experiment teaches us what a causal factor (in this case, gravity) does on its own, when no disturbances (such as air resistance) are present. To assume that businesses are located on a straight line of length *l*, by contrast, does not teach us what the other causal factors (transportation costs, profit maximisation, inelastic demand, etc.) do when geography is absent.

The problem with non-Galilean assumptions is that they make the model result specific to the situation that is being modelled. There is no way to tell from just inspecting the model that it is one subset of assumptions that is driving the result rather than another (cf. Cartwright 1999). And therefore we do not know where to look for 'truth in the model': all we know is that the model result depends on all of a model's assumptions and that many of the model's assumptions are false of any empirical situation we might wish to explain.

It is of course the case that in principle one can test a model result for robustness. Thus, in principle, we can determine which model assumptions drive a result, and from which assumptions results are to some extent independent. Some have even claimed that conducting robustness tests constitutes a significant part of economic practice (Kuorikoski et al. 2010). Indeed, many economic papers contain a section in which robustness is given *some* consideration. But by and large, robustness tests are not possible, and if possible and performed, their result is negative.

Hotelling's model is once more a case in point. The last two pages of his article concern modifications of the original model. Not a single calculation is made, all 'extensions' appear to be based on guesswork. And there is a reason: robustness tests are very hard to perform, and not infrequently impossible, because the mathematics does not allow it altogether or is too difficult for the researcher at hand. About relaxing the inelastic demand assumption, for instance, Hotelling says (1929, p. 56):

> The problem of the two merchants on a linear market might be varied by supposing that each consumer buys an amount of the commodity in question which depends on the delivered price. If one tries a particular demand function the mathematical complications will now be considerable, but for the most general problems elasticity must be assumed.

A paragraph below that he asserts without proof: 'With elastic demand the observations we have made on the solution will still for the most part be qualitatively true … '.

This turned out not to be the case – unless what we take as Hotelling's 'observations' is broad enough to include minimum differentiation, maximum differentiation and everything in between. A recent survey article summarises the following findings regarding changes in the elasticity assumption (Brenner 2001, pp. 14–15):

> The study of Hinloopen and Marrewijk (1999) examines a similar setup where transport costs are linear and the reservation price is constant across consumers. Given the reservation price is sufficiently high, the original Hotelling result holds in which no price equilibrium exists. If the reservation price is low, firms become local monopolists which leads to a continuum of equilibrium locations including maximum and intermediate differentiation. However, reservation prices in-between imply symmetric equilibrium locations where the distance between the firms is between one fourth and one half of the market. For a range of reservation prices there exists a negative relationship between this value and the amount of differentiation. Thus, summarizing we conclude that given a price equilibrium exists for

> the duopoly Hotelling model with uniformly distributed consumer preferences on the unit interval then the higher the elasticity of demand the less firms will differentiate.

The second sentence requires some comment. It is somewhat ironic that 50 years after Hotelling published his paper, his main result – that there is 'stability in (price!) competition' – was shown to be incorrect (D'Aspremont, Gabszewicz, and Thisse 1979). However, that was not too damaging for the minimum differentiation principle because there are a variety of settings in which that result holds, including a game theoretic set-up without price competition (e.g., Osborne 2004, sec. 3.3) and one in which products and consumers are sufficiently heterogeneous (De Palma, Ginsburgh, Papageorgiou, and Thisse 1985).

Let me mention just two further modifications of the original Hotelling set-up. When the number of competitors is three rather than two, there is no stability because the one in the middle will always move to either side to regain its market (Lerner and Singer 1937). Finally, the exact functional form for transportation costs matters. D'Aspremont et al. (1979) showed Hotelling's original result to be incorrect but then went on to find an equilibrium in a setting that is as close as possible to Hotelling's. They found one in a setting that is *identical* to Hotelling's with the exception that transportation costs are now quadratic rather than linear – only that in this setting a principle of *maximum* differentiation holds!

With robustness out of the window, the distinction Mäki needs between 'strategic falsehoods' introduced to isolate a causal power or mechanism of interest and the true descriptions of that causal power or mechanism cannot be sustained. The model result depends on the entire array of assumptions. Consequently, if these assumptions are false of an envisaged target system, we cannot expect the causal power or mechanism to operate in the target system. This is detrimental to our explanatory endeavour. Suppose we observe an instance of minimum differentiation as in Hotelling's cider or US politics. Does Hotelling's model explain that phenomenon under this reading of models? Not if we know some of the model assumptions to be false of the phenomenon and the result – the explanandum – to be dependent on these assumptions.

4.2 Economic models are not explanatory

A number of economic methodologists have denied that economic models are, by themselves, explanatory. Best known in the field is probably Dan Hausman's account of models as conceptual explorations. On this view, models as such do not make claims about the world. Rather, they define predicates, and modelling can be seen as an exercise in exploring conceptual possibilities. Only in conjunction with a theoretical hypothesis of the form 'target system *T* is like model *M*' does a model say something about the world and as a consequence may be explanatory (Hausman 1992, sec. 5.2).

If models are physical or, more frequently, abstract entities, as has been defended here, Hausman's view that a model can only be informative in conjunction with a theoretical hypothesis or some such is of course correct. A physical thing or mathematical structure is not about anything. It is humans who make a thing into a model of some target system *T* by saying they will use the thing as a model of *T*. This can be done explicitly by specifying a theoretical hypothesis or implicitly by simply using the model. Thus, without agents there is no representation and, *a fortiori*, no explanation.

The problem with Hausman's account from our point of view is that he just shifts the issue from the one about false models to the one about false theories. To him, a model plus a theoretical hypothesis is a theory, and thinking of models as conceptual explorations obviously does not help with the question of how false *theories* can be explanatory. We could simply reformulate everything that has been said so far as a problem not only for

models, but for models plus their associated theoretical hypothesis. Hausman certainly does not hold that models plus hypotheses, to wit, theories are not explanatory.

Anna Alexandrova (2008; Alexandrova and Northcott 2009) has given an account of models as open formulae, which also denies that models as such are explanatory. But she holds the stronger view that models, even including a specification of what they are models of, are not explanatory. Rather, models play a heuristic role in which they suggest causal hypotheses, which then are to be tested in experiments.

More specifically, Alexandrova holds that (2008, p. 396; emphasis in original): 'models function as frameworks for formulating hypotheses, or as *open formulae*'. An open formula is a schematic sentence of the form (Alexandrova 2008 p. 392; footnote suppressed): 'In a situation of type x with some characteristics that may include $\{C_1 \ldots C_n\}$, a certain feature F causes a certain behavior B'. The free variable in the open formula is the x, the model specifies the C's, F's and B's. Thus, for instance, an open formula suggested by an economic model may read thusly: 'In a situation of type x, which includes that values are private and some other conditions (the C's) obtain, first-price auction rules (F) cause bids below true valuation (B)'. To move from model to explanation we have to: (1) identify an open formula on the basis of the model; (2) fill in the x so as to arrive at a causal hypothesis; and (3) confirm the causal hypothesis experimentally (Alexandrova 2008, p. 400).

It is clear, then, that in Alexandrova's account models do not play an explanatory role. Models are heuristic devices that suggest causal hypotheses, which, if experimentally confirmed, may be used in causal explanations of phenomena. But this throws the baby out with the bath water. Thousands of economic models have been adduced to explain real-world phenomena without ever having been tested in the lab or elsewhere. In the context of preparing experiments for policy, models may well serve the heuristic function Alexandrova describes. To be fair, she does not claim more than that. More broadly, however, models are regarded as explanatory by themselves. One may of course deny that they are but then arguments have to be given, and it must be explained why a large part of the economics profession thinks otherwise. The open-formulae account therefore ignores rather than solves the problem.

Till Grüne-Yanoff (2009), finally, holds that models prove modal hypotheses, a view with which I have also toyed in Reiss (2008a). Grüne-Yanoff writes that folk wisdom is full of modal claims such as 'Necessarily, segregation is a consequence of racist preferences' or 'It is impossible that intelligent behaviour be produced without a "vitalistic" element present in the organism' (2009, p. 96). We learn about the world from models such as Schelling's (1978) model of racial segregation or Hull's psychic machines, Walter's tortoises and Newell and Simon's simulations, as discussed in Schlimm (2009), that there are possible worlds in which the held beliefs are not true. We learn a possibility result that racial segregation can result from non-racist preferences; that machines can produce intelligent behaviour. (See also Reiss 2008a where I have discussed this function of economic models in some detail.)

Possibility hypotheses, as much as they might teach us about the world, do not explain economic phenomena. It may have been an enormously valuable insight that racial segregation does not have to be the result of outright racism. But this hypothesis at best shows us that actual segregation *can* result otherwise, not that it does so, even in a single case. Economic models therefore may well play the role Grüne-Yanoff ascribes to them, but that they do so does advance our quest on why economic models explain.

All these views, then, ignore rather than solve the problem. Clearly, economic models perform functions other than that of providing explanations. Conceptual explorations (Hausman), heuristics for constructing hypotheses (Alexandrova) and establishing modal

hypotheses (Grüne-Yanoff, Reiss) are salient non-explanatory functions of models, and there may well be others. But some models also explain, and it is this function that the views discussed in this subsection cannot account for.

4.3 Explanation does not require truth

The last premiss of our explanation paradox was that genuine explanation requires truth. This is a widely held belief among philosophers. For most of its history, it was a condition on the logical positivists' DN-model of scientific explanation. It is, necessarily in my view, a condition on acceptable causal explanations. It is very intuitive: telling stories or out-and-out lies is not giving explanations. The truth may not explain much, but without giving at least a slice of truth we have not explained anything. Or so it seems.

When it became apparent that the DN-model of explanation is likely to be irretrievably flawed, philosophers of science sought alternatives, some of which made do without the truth requirement. I will discuss such an account of scientific explanation in detail momentarily. First, however, let us examine one final view of models, given by a prominent economic theorist, experimentalist and methodologist.

Robert Sugden subscribes to the first two premisses of our paradox, as indicated by the following statements (Sugden 2009, p. 3):

> Economic theorists construct highly abstract models. If interpreted as representations of the real world, these models appear absurdly unrealistic; yet economists claim to find them useful in understanding real economic phenomena.

More recent work confirms that Sugden thinks that economic models can be explanatory (Sugden 2011, especially p. 733). To dissolve the paradox, then, he must reject its third premiss. He does so by proposing an account of models as 'credible worlds' (Sugden 2011, p. 2000). A credible world is a deliberate construction, by the modeller, of an abstract entity: a parallel or counterfactual world that, to a greater or lesser extent, resembles aspects of our own world. To learn about the latter, inductive inferences analogous to those from one instance of a type to another are needed. Thus, what we learn from studying Baltimore, Philadelphia, New York, Detroit, Toledo, Buffalo and Pittsburgh we infer to be true in Cleveland as well (Sugden 2000, p. 24). Analogously, we may infer a model result to hold true of a real-world phenomenon. But, according to Sugden, we do so only to the extent that the parallel world depicted by the model is 'credible'. Credibility is thus a key notion in this account. Sugden explains (2009, p. 18; emphasis in original):

> We perceive a model world as credible by being able to think of it as a world that *could* be real – not in the sense of assigning positive subjective probability to the event that it *is* real, but in the sense that it is compatible with what we know, or think we know, about the general laws governing events in the real world.

Sugden explicitly rejects hypothesis three of our paradox (Sugden 2009; emphasis in original): 'Credibility is not the same thing as truth; it is closer to *verisimilitude* or *truthlikeness*'. There is neither the space nor the need here for rehearsing the notorious problems with verisimilitude (for an attempt to cash out the notion in an economics context, see Niiniluoto 2002). What we have to do instead is to consider whether 'credibility' can act as a stand-in for explanatoriness.

I want to address this question at two levels: a descriptive level, which considers whether practising economists hold that (only) credible models are explanatory; and a prescriptive level, which considers whether economists have good reason to do so. Anyone

familiar with the way modelling proceeds in economics will agree that Sugden's account is largely, descriptively adequate. 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. And most economists definitely consider good models explanatory (these are all examples discussed by Sugden; emphases added):[5]

> The example of used cars captures the essence of the problem. From time to time one hears either mention of or surprise at the large price difference between new cars and those which have just left the showroom. The usual lunch table justification for this phenomenon is the pure joy of owning a 'new' car. We offer a different *explanation*. (Akerlof 1970, p. 489)

> Models tend to be useful when they are simultaneously simple enough to fit a variety of behaviors and complex enough to fit behaviors that need the help of an *explanatory model*. (Schelling 1978, p. 89)

> A different *explanation* of herd behavior, which, like the present work is based on informational asymmetries, was suggested in an interesting recent paper by Scharfstein and Stein [1990]. The key difference between their *explanation* and the one suggested here is that their *explanation* is based on an agency problem; in their model the agents get rewards for convincing a principal that they are right. This distortion in incentives plays an important role in generating herd behavior in their model. By contrast, in our model agents capture all of the returns generated by their choice so that there is no distortion in incentives. (Banerjee 1992)

However, we need to ask whether the fact that an economist (or the economics community) regards a model as credible is also a good reason for them to hold that it genuinely explains. Here is where I disagree with Sugden: the 'credibility' of an account of a phenomenon of interest to an individual or a group of researchers is not *per se* a reason to accept it as an explanation of the phenomenon.[6] Many factors affect judgements of credibility, most of which have no essential relationship with explanatoriness: the specific experiences and values of an individual, his or her upbringing and educational background, local customs and culture, social norms and etiquettes of a community of researchers, its theoretical preferences and history.

In Reiss 2008a I argued that economists' subjective judgements of plausibility or credibility are strongly influenced by their theoretical preferences for models that are mathematised, employ the tools of rational choice theory and solve problems using equilibrium concepts. Such preferences are no good reason for considering models with these characteristics as explanatory. Additional arguments would have to be given. One could hold with Galileo, for instance, that the world is Pythagorean, that the book of nature is written in the language of mathematics. (And then go on to argue that therefore genuine explanations have to be couched in mathematical language.) Many readers today will regard this particular claim as highly implausible, but an argument based on a claim *of that kind* is needed.

Resources for such an argument can be found in conceptions of scientific explanation that compete with the causal model. In particular, the view of explanation of successful unification of diverse phenomena is fruitful in this context. Let us examine in some detail Philip Kitcher's (1981) account of explanation (omitting some of its technical details for brevity) and then see whether we can supplement Sugden's proposal in such a way as to resolve our paradox.

Central to Kitcher's account is the notion of an *argument pattern*. Kitcher defines a general argument pattern as consisting of the following: a schematic argument, a set of sets of filling instructions containing one set of filling instructions for each term of the schematic argument, and a classification for the schematic argument (1981, p. 516).

A schematic argument is an argument in which some of the non-logical terms have been replaced by dummy letters; filling instructions are directions specifying how to substitute dummy letters with terms such that sentences obtained are meaningful within the theory; a classification determines which sentences are premisses and which are conclusions, and what rule of inference to use.

Argument patterns are meant to be stringent in that schematic sentences, filling instructions and classification jointly restrict the number of arguments that can be recovered from an argument pattern. An argument pattern that allows the generation of any argument is uninteresting from a scientific point of view. (As an aside: this is a version of Karl Popper's idea that the more conceivable empirical phenomena a scientific theory *excludes*, does *not* predict, the better the theory.)

Suppose K is the set of sentences accepted by some scientific community. A set of arguments that derives some members of K from other members is a *systematisation* of K. Recall that argument patterns can be used to generate arguments. Kitcher calls sets of argument patterns such that every argument in a systematisation of K is an instantiation of a pattern in that set a *generating set*. The *basis* of systematisation is a generating set that is complete (in that every argument which is acceptable relative to K and which instantiates a pattern belongs to the systematisation) and has the greatest unifying power. Finally, the unifying power of a basis (with respect to K) varies directly with: (1) the number of conclusions that can be derived from the set of arguments it generates; and (2) the stringency of its argument patterns, and it varies inversely with the number of its members (Kitcher 1981, p. 520).

Intuitively, the more conclusions that can be derived from using the same set of argument patterns again and again, the more stringent the argument patterns; and the smaller the set of argument patterns needed to derive the conclusions, the greater the unifying power of a basis. While the precise formulation is certainly something one can argue about,[7] Kitcher's notion of unifying power expresses a desideratum many economists require of a good explanation.[8] Indeed, Milton Friedman, while of course avoiding the term explanation, makes the point explicitly. He says that a good theory is at the same time simple and fruitful, and explains (Friedman 1953, p. 10):

> A theory is 'simpler' the less the initial knowledge needed to make a prediction within a given field of phenomena; it is more 'fruitful' the more precise the resulting prediction, the wider the area within which the theory yields predictions, and the more additional lines for further research it suggests.

Let us now address the lacuna Sugden's account of models left. In his view, a model that describes a world that is 'credible' is one that is explanatory. Above I have argued that the credibility of a model does not lend it explanatory power in itself. But what if economists regard models that are unifying as particularly credible? This would allow us to make sense of their demand to make models mathematical and use the principles of rational choice theory and equilibrium concepts – all these form part of argument patterns from which descriptions of a large range of empirical phenomena can be derived. A credible model is one that is explanatory *because* it is unifying.

Why might unifying power be that which lends explanatoriness to a model? Why do we believe that a set of argument patterns that allows us to derive descriptions of a larger range of phenomena of interest is one that is more explanatory? Because to no small extent it is the business of science to achieve cognitive economy – or at least this is one way of thinking about what science tries to achieve. A social practice that told a different story about every phenomenon we ask questions about would not be called a science because it

would not *systematise* what we know about the world. It would not *reduce the number of brute facts* we need to know to understand phenomena of interest. It would not enable us to use what we know to make inferences about new, hitherto unobserved phenomena. Unifying power is surely not the only thing we might seek when we seek explanations of economic phenomena. But the idea that accounts are explanatory to the extent that they are unifying is defensible.

It is unfortunate, therefore, that the argument patterns economics tends to produce are at best spuriously unifying. This is to say that *look very much like* having been defined by a set of assumptions that could be instantiations of generating sets with high unifying power. But in fact they are not.

The problem lies with the notion of stringency. Recall that the unifying power of an argument pattern varies directly with its stringency and that the more arguments a pattern *disallows* to be recovered from it (by specifying highly restrictive schematic sentences, filling instructions and classifications), the more stringent it is. The argument patterns of economics are not at all stringent.

A notorious example is a schematic sentence such as 'Consumers act so as to maximise utility' or, to use a dummy letter in place of a non-logical term, 'Consumers act so as to maximise *U*'. What are the restrictions on the filling instructions for *U*? Answer: very few if any. Often enough people are modelled as deriving utility from some material gain but models do not cease to be economics models if they are more interested in immaterial goods such as reputation or fame, or world peace for that matter. The same is of course true of producers who are said to maximise 'profits'. 'Profits' may be monetary but often they are not. Hotelling-like settings have often been used, for example, to model electoral competition (see for instance Osborne 2004, sec. 3.3). Political parties, of course, maximise votes, not profits.

What make matters worse is that not only 'utility' can be replaced by a dummy letter but also by a 'consumer' (or 'producer'), with similarly unrestrictive filling instructions. In particular, Don Ross has been arguing that not only an 'economic agent' does not have to be a human person, but also that the economics formalism is more likely to work in other species or at sub or super human scales (e.g., Ross 2009). People's preferences are sometimes time inconsistent because of hyperbolic discounting, in apparent violation of economic theory. But, argues Ross, the mistake does not lie with economic theory. Rather, we should move beyond anthropocentric neoclassicism and cease to think of persons as the necessary bearers of economic agency. Consistency can be restored, for instance, by conceiving of human behaviour as the result of a bargaining process between various subpersonal economic agents such as 'short-term' and 'long-term interests' (see also Ross 2005, especially chap. 8).

And the maximisation principle is one of the few principles worth mentioning in economic theory. Schematic sentences and filling instructions thus do not restrict the range of arguments that can very much be generated. The same is true of the classification. The classification contains rules of inference. The most important inference rule in economics is 'Solve the model using an equilibrium concept'. But of course there are many equilibrium concepts. Especially in game theory there is an abundance: the (pure strategy/mixed strategy) Nash equilibrium and its various refinements: subgame perfect equilibrium, trembling-hands equilibrium, Markov-perfect equilibrium, sequential equilibrium, perfect Bayesian equilibrium, evolutionary stable equilibrium and so on (see for instance Fudenberg and Tirole 1991). Equilibrium concepts do not restrict the range of arguments that can be generated much either. To be sure, the inference rule 'use a Nash equilibrium to solve a game' *does* restrict the solution space. But given that: (1) there

are normally many Nash equilibria; (2) there are no clear rules which among a set of Nash equilibria to select; (3) from the point of view of economic theory the justification for using the Nash equilibrium as solution concept is very thin, it would be hard to maintain that economic theory greatly restricts the number of arguments that can be generated from an argument pattern (see Reiss forthcoming, chap. 4).

Friedman (1953) wrote that a theory worth its name consists of a language and a body of substantive hypotheses. Contemporary economic theory does well on the language (the language of logic and mathematics, of rational choice theory and of equilibrium concepts) but lacks substantive hypotheses – schematic sentences that have genuine content in that they (in conjunction with filling instructions and a classification) restrict the number of sentences that can be generated from them. To claim that contemporary economic theory is unifying is therefore like saying to express economic ideas in Italian is unifying. Whatever economists think when they say they provide explanations of this or that phenomenon, the accounts they give are not explanatory qua the unifying power of the argument patterns from which they are derived.

5. Conclusion

The curious thing about genuine paradoxes is that they are not so easily resolved. True, one can always resolve a paradox by fiat – by rejecting one or more of the hypotheses that make up the paradox – but this usually means to ignore the problem. Moreover, it creates the need to explain why so many people believe the claim if it is so unmistakably and so recognisably wrong.

The paradox of economic modelling is genuine in this sense. I think that previous attempts to resolve it have failed, and I do not see many likely avenues for future attempts. Perhaps, thinking about how models explain in ways different from the usual causal and unificationist paradigms is a way forward.[9] But before such a new way of thinking about explanation is forthcoming and shown to fit contemporary economic modelling, the rational response to the paradox is to remain baffled.

Acknowledgements

I thank Nancy Cartwright, Till Grüne-Yanoff, David Teira and two anonymous referees for JEM and the EIPE Reading Group Section 'C' for their invaluable comments. Special thanks to John Davis and Wade Hands for their incredibly efficient editorship. Financial support of the City of Paris (through a grant 'Research in Paris') and the Spanish Government (research projects FFI2008-01580/CONSOLIDER INGENIO CSD2009-0056) is gratefully acknowledged.

Notes

1. There is a vast literature on the causal model of explanation in philosophy of science in general and in economic methodology in particular, far more than I can reasonably cite here. Some core texts are: Salmon (1984), Woodward (2003) on causal explanation in general; and Little (1991, 1998); Kincaid (1996, 1997) and Elster (2007) on causal/mechanical explanation in the social sciences.
2. Cournot was assuming two producers of an identical good who maximise profits by setting quantities at a given price. Bertrand objected that in such a setting either producer would have an incentive to ever so slightly reduce the price he charges, thereby taking away all his competitor's business and nearly doubling his profits. The other will respond by still lowering the price. He argued that by only using quantities as independent variables instead of prices the fallacy can be contained (see Hotelling 1929, p. 43).
3. I omit Wimsatt's type-6 and 7 idealisations because they are not relevant in the context of explanation.

4. See also Strevens (2007) who argues more generally that idealisations serve to tell us what is not causally relevant to the outcome of interest.
5. Milton Friedman is an exception to this rule, at least according to his explicit methodological statement in his 1953 work. For Friedman, the issue is one about the correct understanding of the aims of economics, and to him explanation was not among them. He did *not* reject the idea that credibility is a good indicator or constitutive of explanatoriness.
6. To be fair I should mention that Sugden does not put it this way. But it is an implication of views he does hold explicitly that economic models aim to explain real-world phenomena (e.g., 'Similarly, if the theorist is offering a tool that is intended to be used in explaining real-world phenomena, a convincing demonstration must display the tool explaining something', (Sugden 2009, p. 25), and that judgements of credibility determine if they do so (e.g., 'I argued that the gap between model world and real world has to be crossed by inductive inference, and that inductive inference depends on subjective judgements of "similarity", "salience" and "credibility"' (Sugden 2009, p. 4).
7. Kitcher himself points out, for instance, that this way of putting it is too simple as members of sets of argument patterns may be similar to each other (in that they use the same core), and similarity among the argument patterns increases its unifying power (Friedman 1953, p. 521). I ignore this detail as all I require here is the intuitive idea and some detail regarding the notion of *stringency*.
8. I am not the only one to have made this observation, see for instance Mäki (2001) and Lehtinen and Kuorikoski (2007). I wrote about it first in Reiss 2002.
9. To my knowledge, the only serious attempt has been made by Bokulich (2009). Her account of model explanation, however, does not seem to fit economics very well, for reasons that are very similar to those given in response to Kitcher's account.

References

Akerlof, G. (1970), 'The Market for 'Lemons': Quality Uncertainty and the Market Mechanism', *Quarterly Journal of Economics*, 84(3), 488–500.

Alexandrova, A. (2008), 'Making Models Count', *Philosophy of Science*, 75(3), 383–404.

Alexandrova, A., and Northcott, R. (2009), 'Progress in Economics: Lessons From the Spectrum Auctions', in *The Oxford Handbook of Philosophy of Economics*, eds. H. Kincaid and D. Ross, Oxford, NY: Oxford University Press, pp. 306–336.

Banerjee, A. (1992), 'A Simple Model of Herd Behavior', *Quarterly Journal of Economics*, 107(3), 797–817.

Bokulich, A. (2009), 'How Scientific Models Can Explain', *Synthese*, 180(1), 33–45.

Box, G., and Draper, N. (1987), *Empirical Model-Building and Response Surfaces*, New York: John Wiley & Sons.

Brenner, S. (2001), *Determinants of Product Differentiation: A Survey*, Berlin: Humbold University.

Cartwright, N. (1983), *How the Laws of Physics Lie*, Oxford, NY: Oxford University Press.

——— (1989), *Nature's Capacities and Their Measurement*, Oxford, NY: Clarendon.

——— (1999), *The Vanity of Rigor in Economics: Theoretical Models and Galileian Experiments* (CPNSS Discussion Papers DP 43/99), London: Centre for Philosophy of Natural and Social Sciences.

D'Aspremont, C., Gabszewicz, J.J., and Thisse, J.F. (1979), 'On Hotelling's "Stability in Competition"', *Econometrica*, 47(5), 1145–1150.

De Palma, A., Ginsburgh, V., Papageorgiou, Y.Y., and Thisse, J.F. (1985), 'The Principle of Minimum Differentiation Holds Under Sufficient Heterogeneity', *Econometrica*, 53(4), 245–252.

Elster, J. (2007), *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*, Cambridge, NY: Cambridge University Press.

Friedman, M. (1953), 'The Methodology of Positive Economics', in *Essays in Positive Economics*, Chicago, IL: University of Chicago Press, pp. 3–46.

Frigg, R., and Hartmann, S. (2006), 'Models in Science', in *Stanford Encyclopedia of Philosophy*, ed. E. Zalta, http://plato.stanford.edu/entries/models-science/

Fudenberg, D., and Tirole, J. (1991), *Game Theory*, Cambridge, MA: MIT Press.

Grüne-Yanoff, T. (2009), 'Learning from Minimal Economic Models', *Erkenntnis*, 70(1), 81–99.

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

Hinloopen, J. and Marrewijk, C.V. (1999), 'On the limits and possibilities of the principle of minimum differentiation', *International Journal of Industrial Organization*, 17, 735–750.
Hotelling, H. (1929), 'Stability in Competition', *Economic Journal*, 39(153), 41–57.
Kincaid, H. (1996), *Philosophical Foundations of the Social Sciences*, New York: Cambridge University Press.
——— (1997), *Individualism and the Unity of Science*, Lanham, MD: Rowman and Littlefield.
Kitcher, P. (1981), 'Explanatory Unification', *Philosophy of Science*, 48, 507–531.
Kuorikoski, J., Lehtinen, A., and Marchionni, C. (2010), 'Economic Modelling as Robustness Analysis', *British Journal for the Philosophy of Science*, 61(3), 541–567.
Lehtinen, A., and Kuorikoski, J. (2007), 'Computing the Perfect Model: Why Do Economists Shun Simulation?', *Philosophy of Science*, 74, 304–329.
Lerner, A., and Singer, H. (1937), 'Some Notes on Duopoly and Spatial Competition', *Journal of Political Economy*, 45, 145–186.
Little, D. (1991), *Varieties of Social Explanation*, Boulder, CO: Westview Press.
——— (1998), *Microfoundations, Method, and Causation: On the Philosophy of the Social Sciences*, New Brunswick, NJ: Transaction Publishers.
Mäki, U. (1992), 'On the Method of Isolation in Economics', in *Idealization IV: Structuralism, Intelligibility in Science*, ed. C. Dilworth, Amsterdam/Atlanta, GA: Rodopi, pp. 317–351.
——— (1994), 'Isolation, Idealization and Truth in Economics', in *Idealization in Economics*, eds. B. Hamminga, N. de Marchi, Amsterdam: Rodopi, pp. 147–168.
——— (2001), 'Explanatory Unification: Double and Doubtful', *Philosophy of the Social Sciences*, 31, 488–506.
——— (2005), 'Models Are Experiments, Experiments Are Models', *Journal of Economic Methodology*, 12, 303–315.
——— (2009), 'Missing the World: Models as Isolations and Credible Surrogate Systems', *Erkenntnis*, 70(1), 29–43.
——— (2011), 'Models and the Locus of Their Truth', *Synthese*, 180, 47–63.
McAllister, J. (2004), 'Thought Experiments and the Belief in Phenomena', *Philosophy of Science*, 71(5), 1164–1175.
McMullin, E. (1985), 'Galileian Idealization', *Studies in the History and Philosophy of Science*, 16, 247–273.
Niiniluoto, I. (2002), 'Truthlikeliness and Economic Theories', in *Fact and Fiction in Economics: Models, Realism and Social Construction*, ed. U. Mäki, Cambridge: Cambridge University Press, pp. 214–228.
Osborne, M. (2004), *An Introduction to Game Theory*, Oxford: Oxford University Press.
Reiss, J. (2002), 'Epistemic Virtues and Concept Formation in Economics', unpublished Ph.D. dissertation, London, London School of Economics.
——— (2008a), *Error in Economics: Towards a More Evidence-Based Methodology*, London: Routledge.
——— (2008b), 'Social Capacities', in *Nancy Cartwright's Philosophy of Science*, eds. S. Hartman and L. Bovens, London: Routledge, pp. 265–288.
——— (forthcoming), *The Philosophy of Economics*, New York: Routledge.
Ross, D. (2005), *Economic Theory and Cognitive Science: Microexplanation*, Cambridge, MA: MIT Press, pp. 245–279.
——— (2009), 'Integrating the Dynamics of Multi-Scale Economic Agency', in *Oxford Handbook of Philosophy of Economics*, eds. H. Kincaid and D. Ross, Oxford, NY: Oxford University Press.
Salmon, W. (1984), *Scientific Explanation and the Causal Structure of the World*, Princeton, NJ: Princeton University Press.
Scharfstein, D., and Stein, J. (1990), 'Herd Behavior and Investment', *American Economic Review*, 80(3), 465–479.
Schelling, T. (1978), *Micromotives and Macrobehavior*, New York: Norton.
Schlimm, D. (2009), 'Learning from the Existence of Models: On Psychic Machines, Tortoises, and Computer Simulations', *Synthese*, 169(3), 521–538.
Strevens, M. (2007), *Why Explanations Lie: Idealization in Explanation*, New York: New York University.
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(1), 3–27.

——— (2011), 'Explanations in Search of Observations', *Biology and Philosophy*, 26, 717–736.
Taylor, P.J. (1985), *Construction and turnover of multi-species communities: a critique of approaches to ecological complexity*. Ph.D. dissertation, Department of Organismal and Evolutionary Biology, Harvard University.
Wimsatt, W. (2007), *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*, Cambridge, MA: Harvard U niversity Press.
Woodward, J. (2003), *Making Things Happen*, Oxford, NY: Oxford University Press.