TWENTY-SIX

Credible Worlds

The Status of Theoretical Models in Economics

Robert Sugden

Robert Sugden (1949–) is a professor of economics at the University of East Anglia. His research uses theoretical, experimental, and philosophical methods to investigate issues in welfare economics, choice under uncertainty, pro-social behavior, the emergence of conventions and norms, economic methodology, and philosophical economics. He is probably best known for developing "regret theory" (with Graham Loomes) and for *The Economics of Rights, Cooperation and Welfare*, one of the first applications of evolutionary game theory to social theory and moral philosophy. Currently, he holds a research fellowship from the Economic and Social Research Council for work on reconciling behavioral and normative economics.

Overview

Using as examples Akerlof's 'market for "lemons"' and Schelling's 'checkerboard' model of racial segregation, this paper asks how economists' abstract theoretical models can explain features of the real world. It argues that such models are not abstractions from, or simplifications of, the real world. They describe counterfactual worlds which the modeller has constructed. The gap between model world and real world can be filled only by inductive inference, and we can have more confidence in such inferences, the more credible the model is as an account of what could have been true.

1. Introduction

I write this paper not as a methodologist or as a philosopher of social science – neither of which I can make any claim to be – but as a theoretical economist. I have spent a considerable part of my life building economic models, and examining the models that other economists have built. I believe that I am

Journal of Economic Methodology, vol. 7 (2000): 1–31. Reprinted by permission of Taylor and Francis, Ltd.

making reasonably good use of my talents in an attempt to understand the social world. I have no fellow-feeling with those economic theorists who, off the record at seminars and conferences, admit that they are only playing a game with other theorists. If their models are not intended seriously, I want to say (and do say when I feel sufficiently combative), why do they expect me to spend my time listening to their expositions? Count me out of the game. At the back of my mind, however, there is a trace of self-doubt. Do the sort of models that I try to build really help us to understand the world? Or am I too just playing a game, without being self-critical enough to admit it?

My starting point is that model-building in economics has serious intent only if it is ultimately directed towards telling us something about the real world. In using the expression 'the real world' – as I shall throughout the paper – I immediately reveal myself as an economic theorist. This expression is standardly used by economic theorists to mark the distinction between the world inside a model and the 'real' world outside it. Theory becomes just a game when theorists work entirely in the world of models. As an analogy, we might think of chess, which was once a model of warfare, but has become a game – a self-contained world with no reference to anything outside itself.

My strategy is to focus on two models – George Akerlof's 'market for lemons', and Thomas Schelling's 'checkerboard city' – which exemplify the kind of model-building to which I aspire. Of course, these are not typical examples of economic models: they represent theory at its best. Nevertheless, at least at first sight, these models have many of the vices that critics attribute to theoretical economics: they are abstract and unrealistic and they lead to no clearly testable hypotheses. It would be easy to caricature them as examples – perhaps unusually imaginative and, from a mathematical point of view, unusually informal examples – of the games that economic theorists play. Thus, they provide suitable case studies for an attempted defence of model-building in economics.

I believe that each of these models tells us something important and true about the real world. My object is to discover just what these models do tell us about the world, and how they do it.

2. Akerlof and the Market for 'Lemons'

Akerlof's 1970 paper 'The market for "lemons"' is one of the best-known papers in theoretical economics. It is generally seen as having introduced to economics the concept of asymmetric information, and in doing so, sparking off what is now a whole branch of economics: the economics of information.

It is a theoretical paper that almost all economists, however untheoretical they might be, would now recognize as important. It is also a paper that just about every economic theorist would love to have written. Because there is no dispute about its value, Akerlof's paper is particularly suitable for my purposes. Everyone can see that this is a major contribution to economics.[1] The puzzle is to say exactly what the contribution is. Is Akerlof telling us anything about the real world, and if so, what?

It is worth looking closely at the structure of the paper. Here is the opening paragraph:

> This paper relates quality and uncertainty. The existence of goods of many grades poses interesting and important problems for the theory of markets. On the one hand, the interaction of quality differences and uncertainty may explain important institutions of the labour market. On the other hand, this paper presents a struggling attempt to give structure to the statement: 'business in underdeveloped countries is difficult'; in particular, a structure is given for determining the economic costs of dishonesty. Additional applications of the theory include comments on the structure of money markets, on the notion of 'insurability', on the liquidity of durables, and on brand-name goods. (Akerlof 1970: 488)

Clearly, Akerlof is claiming that his paper has something to say about an astonishingly wide range of phenomena in the real world. The paper, we are promised, is going to tell us something about the institutions of the labour market, about business in underdeveloped countries, about insurability, and so on. But what kind of thing is it going to tell us? On this point, Akerlof is rather coy. In the case of the labour market, he seems to be promising to explain some features of the real world. (Or is he? See later.) But in the case of business in underdeveloped countries, he is only going to *give structure to a statement* that is often made about the real world. Here, the implication seems to be that Akerlof's model will somehow reformulate an empirical proposition which is generally believed to be true (but might actually be false). In the other cases we are promised comments which are to be understood as applications of the theory he is to present.

Akerlof then says that, although his theory has these very general applications, he will focus on the market for used cars:

> The automobile market is used as a finger exercise to illustrate and develop these thoughts. It should be emphasized that this market is chosen for its concreteness and ease in understanding rather than for its importance or realism. (Akerlof 1970: 489)

On first reading, it is tempting to interpret 'the automobile market' as the market in which real people buy and sell real cars, and to think that Akerlof is going to present some kind of case study. One can see why he might focus on one particular market which is easy to understand, even if that market is not very important on the scale of the economy as a whole. But then what does Akerlof mean when he says that this market is not *realistic*? The object of a case study may be unrepresentative, but it cannot be unrealistic. To make sense of this passage, I think, we have to recognize that it marks a transition between the real world and the world of models. Akerlof is using the real automobile market as an example. But what he is going to present is not an empirical case study; it is a model of the automobile market. Although it is the real market which may be unimportant, it is the model which may be unrealistic.

Akerlof moves straight on to the central section of his paper, section II, entitled 'The Model with Automobiles as an Example'. The transition from reality to model is made again at the very beginning of this section:

> 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. Suppose (for the sake of clarity rather than realism) that there are just four kinds of cars. There are new cars and used cars. There are good cars and bad cars . . . (Akerlof 1970: 489)

The first four sentences are about an observed property of the real world: there is a large price difference between new cars and almost-new ones. Akerlof suggests that, at least from the viewpoint of the lunch table, this observation is difficult to explain. If we assume that Akerlof takes lunch with other economists, the implication is that economics cannot easily explain it; the 'pure joy' hypothesis sounds like an *ad hoc* stratagem to rescue conventional price theory. So far, then, the mode of argument might be Popperian: there is a received theory which makes certain predictions about market prices; observations of the used car market are contrary to those predictions; therefore, a new theory is needed.[2]

But from the word 'suppose' in the passage above, we move out of the real world and into the world of the model. Akerlof sets up an imaginary world; he makes no pretence to describe any real market. In this world, there are two groups of traders, 'type one' and 'type two'. All traders of a given type are alike. There are *n* cars, which differ only in 'quality'. Quality is measured in money units and is uniformly distributed over some range. Each group of traders maximizes an aggregate utility function. For group one, utility is the sum of the qualities of the cars it owns and the monetary value of its consumption of other goods. For group two, the utility function is the same, except that quality is multiplied by 3/2. Thus, for any given quality of car, the monetary value of a car to type one traders is less than its monetary value to type two traders. All cars are initially owned by type one traders. The quality of cars has a uniform distribution. The quality of each car is known only to its owner, but the average quality of all traded cars is known to everyone.

Akerlof admits that these assumptions are not realistic: they are not even close approximations to properties of the real used-car market. He justifies them as simplifications which allow him to focus on those features of the real market that he wishes to analyse. For example, he defends his assumptions about utility (which implicitly impose risk neutrality) against what he takes to be the more realistic alternative assumption of risk aversion by saying that he does not want to get 'needlessly mired in algebraic complication': 'The use of linear utility allows a focus on the effects of asymmetry of information; with a concave utility function we would have to deal with the usual risk-variance effects of uncertainty and the special effects we have to deal with here' (pp. 490–491).

Akerlof investigates what happens in his model world. The main conclusion is simple and startling. He shows that if cars are to be traded at all, there must be a single market price p. Then:

> However, with any price p, average quality is p/2 and therefore at no price will any trade take place at all: in spite of the fact that *at any given price* [between certain limits] there are traders of type one who are willing to sell their automobiles at a price which traders of type two are willing to pay. (Akerlof 1970: 491)

Finally, Akerlof shows what would happen in the same market if information were symmetric – that is, if neither buyers nor sellers knew the quality of individual cars, but both knew the probability distribution of quality. In this case, there is a market-clearing equilibrium price, and trade takes place, just as the standard theory of markets would lead us to expect. Akerlof ends section II at this point, so let us take stock.

What we have been shown is that in a highly unrealistic model of the used car market, no trade takes place – even though each car is worth less to its owner than it would be to a potential buyer. We have also been given some reason to think that, in generating this result, the crucial property of the model world is that sellers know more than buyers. Notice that, taken literally, Akerlof's result is too strong to fit with the phenomenon he originally promised to explain – the price difference between new and used cars.[3] Presumably, then, Akerlof sees his model as describing in extreme form the workings of some *tendency* which exists in the real used-car market, by virtue of the asymmetry of information which (he claims) is a property of that market. This tendency is a used-car version of Gresham's Law: bad cars drive out good. In the real used-car market, according to Akerlof, this tendency has the effect of reducing the average quality of cars traded, but not eliminating trade altogether; the low quality of traded cars then explains their low price.

Remarkably, Akerlof says nothing more about the *real* market in used cars. In the whole paper, the only empirical statement about the used-car market is the one I have quoted, about lunch-table conversation. Akerlof presents no evidence to support his claim that there is a large price difference between new and almost-new cars. This is perhaps understandable, since he clearly assumes that this price difference is generally known. More surprisingly, he presents no evidence that the owners of nearly-new cars know significantly more about their quality than do potential buyers. And although later in the paper he talks about market institutions which can overcome the problem of asymmetric information, he does not offer any argument, theoretical or empirical, to counter the hypothesis that such institutions exist in the used-car market. But if they do, Akerlof's explanation of price differences is undermined.

However, Akerlof has quite a lot to say about other real markets in section III of the paper, 'Examples and Applications'. In four subsections, entitled 'Insurance', 'The Employment of Minorities', 'The Costs of Dishonesty', and 'Credit Markets in Underdeveloped Countries', Akerlof presents what are effectively brief case studies. We are told that adverse selection in the insurance market is 'strictly analogous to our automobiles case' (p. 493), that 'the Lemons Principle . . . casts light on the employment of minorities' (p. 494), that 'the Lemons model can be used to make some comments on the costs of dishonesty' (p. 495), and that 'credit markets in underdeveloped countries often strongly reflect the Lemons Principle' (p. 497). These discussions are in the style that economists call 'casual empiricism'. They are suggestive, just as the used-car case is, but they cannot be regarded as any kind of test of a hypothesis. In fact, there is no hypothesis. Akerlof never defines the 'lemons principle'; all we can safely infer is that this term refers to the model of the used-car market. Ultimately, then, the claims of section III amount to this: In these four cases, we see markets that are in some way like the model.

The final part of the paper (apart from a very short conclusion) is section IV, 'Countervailing Institutions'. This is a brief discussion, again in the mode of casual empiricism, of some real-world institutions which counteract the problem of asymmetric information. The examples looked at are guarantees, brand names, hotel and restaurant chains, and certification in the labour market (such as the certification of doctors and barbers). The latter example seems to be what Akerlof was referring to in his introduction when he claimed that his approach might 'explain important institutions of the labour market'. Here, the claim seems to be that there are markets which would be like the model of the used-car market, were it not for some special institutional feature; therefore, the model explains those features.

From a Popperian perspective, sections III and IV have all the hallmarks of 'pseudo-science'. Akerlof has not proposed any hypothesis in a form that could be tested against observation. All he has presented is an empirically ill-defined 'lemons principle'. In Section III, he has assembled a fairly random assortment of evidence which appears to confirm that principle. In Section IV, he argues that the real world often is not like the model, but this is to be seen not as refutation but as additional confirmation. What kind of scientific reasoning is this?

3. Schelling's Checkerboard Model of Racial Sorting

My other example of a theoretical model in economics is not quite as famous as the market for lemons, but it is a personal favourite of mine.[4] It also deserves to be recognized as one of the earliest uses of what is now a well-established theoretical method: evolutionary game theory with localized interactions in a spatial structure. This is the chapter 'Sorting and Mixing: Race and Sex' in Schelling's book *Micromotives and Macrobehaviour* (1978).

The book as a whole is concerned with one of the classic themes of economics: the unintended social consequences of uncoordinated individual actions. Using a wide range of novel and surprising examples, Schelling sets out to show that spontaneous human interaction typically generates unintended patterns at the social level; in some cases these patterns are desirable, but in many cases they are not.

Schelling opens this chapter with an extended and informal discussion of segregation by colour and by sex in various social settings. His concern is with patterns of segregation that arise out of the voluntary choices of individuals. One important case of such self-segregation, he suggests, is the housing market of American cities. Blacks and whites[5] tend to live in separate areas; the boundaries of these areas change over time, but the segregation remains. Schelling suggests that it is unlikely that almost all Americans desire to live in such sharply segregated areas. He asks us to consider the possibility that the sharp segregation we observe at the social level is an unintended consequence of individual actions which are motivated only by a preference for not living in an area in which people of the other colour form an overwhelming majority. In the context of tables in a cafeteria for a baseball training camp, Schelling puts his hypothesis like this:

> Players can ignore, accept, or even prefer mixed tables but become uncomfortable or self-conscious, or think that others are uncomfortable or self-conscious, when the mixture is lopsided. Joining a table with blacks and whites is a casual thing, but being the seventh at a table with six players of the opposite colour imposes a threshold of self-consciousness that spoils the easy atmosphere and can lead to complete and sustained separation. (Schelling 1978: 144)

Having discussed a number of cases of self-segregation, both by colour and by sex, and in each case having floated the hypothesis that sharp segregation is an unintended consequence of much milder preferences, Schelling presents a 'self-forming neighbourhood model'. He begins disarmingly: 'Some vivid dynamics can be generated by any reader with a half-hour to spare, a roll of pennies and a roll of dimes, a tabletop, a large sheet of paper, a spirit of scientific enquiry, or, failing that spirit, a fondness for games' (p. 147).

We are instructed to mark out an 8 x 8 grid of squares. The dimes and pennies:

> represent the members of two homogeneous groups – men and women, blacks and whites, French-speaking and English-speaking, officers and enlisted men, students and faculty, surfers and swimmers, the well dressed and the poorly dressed, or any other dichotomy that is exhaustive and recognizable. (Schelling 1978: 147)

We then distribute coins over the squares of the grid. Each square must either be allocated one coin or left empty (it is important to leave some empty spaces). Next, we postulate a condition which determines whether a coin is 'content' with its neighbourhood. For example, we might specify that a coin is content provided that at least one-third of its neighbours (that is, coins on horizontally, vertically or diagonally adjacent squares) are of the same type as itself. Then we look for coins which are not content. Whenever we find such a coin, we move it to the nearest empty square at which it *is* content (even if, in so doing, we make other coins discontented). This continues until there are no discontented coins. Schelling suggests that we try this with different initial distributions of coins and different rules. What we will find, he says, is a very strong tendency for the emergence of sharply segregated distributions of coins, even when the condition for contentedness is quite weak. I have followed Schelling's instructions (with the help of a computer program rather than paper and coins), and I can confirm that he is right. Clearly, Schelling expects that after we have watched the workings of this model, we will find his earlier arguments about real-world segregation more convincing.

The general strategy of Schelling's chapter is remarkably similar to that of Akerlof's paper. Each author is claiming that some regularity R (bad products driving out good, persistent racial segregation with moving geographical boundaries) can be found in economic or social phenomena. Each is also claiming that R can be explained by some set of causal factors F (sellers being better-informed than buyers, a common preference not to be heavily outnumbered by neighbours not of one's own type). Implicitly, each is making three claims: that R occurs (or often occurs); that F operates (or often operates); and that F causes R (or tends to cause it). Neither presents any of these claims as a testable hypothesis, but each offers informal evidence from selected case studies which seems to support the first two claims. Each uses a formal model in support of the claim about causation. In each case, the formal model is a very simple, fully-described and self-contained world. The supposedly causal factors F are built into the specification of the model. In the model world, R is found in an extreme form. This is supposed to make more credible the claim that in the real world, F causes R. But just how is that claim made more credible?

4. Conceptual Exploration

Before going on, we need to consider an alternative reading of Akerlof and Schelling, in which their models are not intended to support any claims about the real world.[6] As Daniel Hausman (1992: 221) has pointed out, theoretical work in economics is often concerned with 'conceptual exploration' rather than 'empirical theorizing'. Conceptual exploration investigates the internal properties of models, without considering the relationship between the world of the model and the real world.

Such work can be seen as valuable, even by someone who insists that the ultimate purpose of model-building is to tell us something about the real world. For example, it can be valuable because it finds simpler formulations of existing theories, or discovers useful theorems within those theories. (Consider Paul Samuelson's demonstration that most of conventional demand theory can be deduced from a few simple axioms about consistent choice.) Or it can be valuable because it discovers previously unsuspected inconsistencies in received theories. (For example, Kenneth Arrow's impossibility theorem can be interpreted as a demonstration of the incoherence of Bergson-Samuelson welfare economics.[7]) There are also instances in which the development of a theory intended for one application has generated results which have later proved to be useful in completely different domains. (Think how much has grown out of John von Neumann and Oskar Morgenstern's exploration of strategies for playing poker.) Thus, to characterize Akerlof's and Schelling's models as conceptual exploration need not be to denigrate them.

So let us consider what we would learn from these models if we interpreted them as conceptual exploration and nothing else. Take Akerlof first. Akerlof's contribution, it might be said, is to show that some implications of the standard behavioural assumptions of economic theory are highly sensitive to the particular simplifying assumptions that are made about knowledge.[8] More specifically, the usual results about Pareto-efficient, market-clearing equilibrium trade can be radically altered if, instead of assuming that buyers and sellers are equally well-informed, we allow some degree of asymmetry of information. The message of Akerlof's paper, then, is that some commonly-invoked theoretical propositions about markets are not as robust as was previously thought. Thus, conclusions derived from models which assume symmetric information should be treated with caution, and new theories need to be developed which take account of the effects of asymmetric information. On this reading, the discussion of used cars is no more than a 'story' attached to a formal model, useful in aiding exposition and comprehension, but which can be dispensed with if necessary.[9] The paper is not about used cars: it is about the theory of markets.

What about Schelling? We might say that Schelling is presenting a critique of a commonly-held view that segregation must be the product either of deliberate public policy or of strongly segregationist preferences. The checkerboard model is a counter-example to these claims: it shows that segregation could arise without either of those factors being present. On this reading, Schelling is making an important contribution to debates about segregation in the real world, but the contribution is conceptual: he is pointing to an error in an existing theory. In terms of the symbols I introduced in section 3, Schelling is not asserting: 'R occurs, F operates, and F causes R'. All he is asserting is: 'R could occur, F could operate, and it could be the case that F caused R'.

It must be said that there is at least some textual evidence that both Akerlof and Schelling are tempted by this kind of interpretation of their models. As I have already suggested, Akerlof often seems to be taking care not to draw inferences about the real world from his model. For example, although he does claim to be offering an explanation of price differences in the real car market, his other references to ‘explanation’ are more nuanced. Notice that in the opening paragraph he does not claim that his model explains important institutions of the labour market: what may (not does) explain them is ‘the interaction of quality differences and uncertainty’. The final sentence of the paper uses a similar formulation: ‘the difficulty of distinguishing good quality from bad . . . may indeed explain many economic institutions’ (p. 500). On one reading of ‘may’ in these passages, Akerlof is engaged only in conceptual exploration: he is considering what sorts of theory are possible, but not whether or not these theories actually explain the phenomena of the real world. However, I shall suggest that a more natural reading is that Akerlof is trying to say something like this: I believe that economists will be able to use the ideas in this paper to construct theories which *do* explain important economic institutions.

Schelling is more explicit about his method, and what it can tell us:

> What can we conclude from an exercise like this? We may at least be able to disprove a few notions that are themselves based on reasoning no more complicated than the checkerboard. Propositions beginning with ‘It stands to reason that . . . ’ can sometimes be discredited by exceedingly simple demonstrations that, though perhaps true, they do not exactly ‘stand to reason’. We can at least persuade ourselves that certain mechanisms could work, and that observable aggregate phenomena could be compatible with types of ‘molecular movement’ that do not closely resemble the aggregate outcomes that they determine. (Schelling 1978: 152)

Schelling does not elaborate on what notions he has disproved. Possibly what he has in mind is the notion that either deliberate policy or the existence of strongly segregationist preferences is a necessary condition for the kind of racial segregation that is observed in American cities. His claim, then, is that he has discredited this notion by means of a counter-example.

Whatever we make of these passages, neither paper, considered as a whole, can satisfactorily be read as conceptual exploration and nothing else. The most obvious objection to this kind of interpretation is that Akerlof and Schelling both devote such a lot of space to the discussion of real-world phenomena. Granted that Akerlof’s treatment of the used car market has some of the hallmarks of a theorist’s ‘story’, what is the point of all the ‘examples and applications’ in his section III, or of the discussion of ‘countervailing institutions’ in section IV, if not to tell us something about how the world really is? This material may be casual empiricism, but it is empiricism none the less. It is not just a way of helping us to understand the internal logic of the model. Similarly, Schelling’s discussion of the baseball training camp is clearly intended as a description of the real world. Its purpose, surely, is to persuade us of the credibility of the hypothesis that real people – it is hinted, people like us – have mildly segregationist preferences. If all we were being offered was a counterexample to a general theoretical claim, such material would be redundant.

Clearly, neither Akerlof nor Schelling wants to claim that his work is a completed theory. The suggestion seems to be that these are preliminary sketches of theories. The models that are presented are perhaps supposed to stand in the sort of relation to a completed theory that a 'concept car' does to a new production model, or that the clothes in a *haute couture* fashion show do to the latest designs in a fashion shop. That is, these models are suggestions about how to set about explaining some phenomenon in the real world. To put this another way, they are sketches of processes which, according to their creators, might explain phenomena we can observe in the real world. But the sense of 'might explain' here is not just the kind of logical possibility that could be discovered by conceptual exploration. (The latter sense could be paraphrased as: 'In principle, it is possible that processes with this particular formal structure could generate regularities with that particular formal structure'.) The theorist is declaring his confidence that his approach is likely to work as an explanation, even if he does not claim so to have explained anything so far.

If Akerlof's and Schelling's disclaimers were to be read as saying 'This work is conceptual exploration and nothing else', they would surely be disingenuous. We are being offered potential explanations of real-world phenomena. We are being encouraged to take these potential explanations seriously – perhaps even to do some of the work necessary to turn these sketches of theories into production models. If we are to do this, it is not enough that we have confidence in the technical feasibility of an internally consistent theory. Of course, having that confidence is important, and we can get it by conceptual exploration of formal models. But what we need in addition is some confidence that the production model is likely to do the job for which it has been designed – that it is likely to explain real-world phenomena. In other words, we need to see a sketch of an *actual* explanation, not just of a logically coherent formal structure. We should expect Akerlof's and Schelling's models to provide explanations, however tentative and imperfect, of regularities in the real world. I shall proceed on the assumption that these models are intended to function as such explanations.

5. Instrumentalism

This brings us back to the problem: How do unrealistic economic models explain real-world phenomena?

Many economists are attracted by the instrumentalist position that a theory should be judged only on its predictive power within the particular domain in which it is intended to be used. According to one version of instrumentalism, the ‘assumptions’ of a theory, properly understood, are no more than a compact notation for summarizing the theory’s predictions; thus, the question of whether assumptions are realistic or unrealistic does not arise. An alternative form of instrumentalism, perhaps more appropriate for economics, accepts that the assumptions of a theory *refer* to things in the real world, but maintains that it does not matter whether those assumptions are true or false. On either account, the assumptions of a theory *function* only as a representation of the theory’s predictions.

Instrumentalist arguments are often used in defence of the neoclassical theory of price determination which assumes utility-maximizing consumers, profit-maximizing firms, and the instantaneous adjustment of prices to market-clearing levels. In the instrumentalist interpretation the object of the neoclassical theory is to predict changes in the prices and total quantities traded of different goods as a result of exogenous changes (such as changes in technology or taxes). On this view, aggregated economic statistics play the same role in economics as the movements of the heavenly bodies through the sky did in early astronomy:[10] they are the only phenomena we want to predict, and the only (or only acceptable) data.[11] The neoclassical theory is just a compact description of a set of predictions. To ask whether its assumptions are realistic is either to make a category mistake (because assumptions do not refer to anything that has real existence) or to miss the point (because, although assumptions refer to real things, the truth or falsity of those references has no bearing on the value of the theory).

But is it possible to understand Akerlof’s and Schelling’s models instrumentally? These models are certainly similar to the neoclassical model of markets in their use of highly simplified assumptions which, if taken literally, are highly unrealistic. But if these models are intended to be read instrumentally, we should expect to find them being used to generate unambiguous predictions about the real world. Further, there should be a clear distinction between assumptions (which either have no truth values at all, or are allowed to be false) and predictions (which are asserted to be true).

In fact, neither Akerlof nor Schelling proposes any explicit and testable hypothesis about the real world. Nor does either theorist maintain an instrumentalist distinction between assumptions and predictions. Akerlof’s case studies seem to be intended as much to persuade us of the credibility of his assumptions about asymmetric information as to persuade us that the volume of trade is sub-optimal. As I have already said, Schelling’s discussion of the baseball camp seems to be intended to persuade us of the credibility of his assumptions about preferences. On the most natural readings, I suggest, Akerlof and Schelling think they are telling us about forces or tendencies which connect *real* causes (asymmetric information, mildly segregationist preferences) to *real* effects (sub-optimal volumes of trade, sharp segregation). Akerlof's and Schelling's unrealistic models are supposed to give support to these claims about real tendencies. Whatever method this is, it is not instrumentalism: it is some form of realism.

6. Metaphor and Caricature

Allan Gibbard and Hal Varian (1978) offer an interpretation of economic models which emphasizes explanation rather than prediction. They characterize a model as the conjunction of two elements: an uninterpreted formal system within which logical deductions can be made, and a 'story' which gives some kind of interpretation of that formal system. With Schelling's checkerboard model apparently in mind, they describe a form of modelling in which the fit of the model to the real world is *casual*:

> The goal of casual application is to explain aspects of the world that can be noticed or conjectured without explicit techniques of measurement. In some cases, an aspect of the world (such as price dispersal, housing segregation, and the like) is noticed, and certain aspects of the micro-situation are thought perhaps to explain it; a model is then constructed to provide the explanation. In other cases, an aspect of the microworld is noticed, and a model is used to investigate the kinds of effects such a factor could be expected to have. (Gibbard and Hal Varian 1978: 672)

This seems a fair description of what both Akerlof and Schelling are doing. But Gibbard and Varian have disappointingly little to say about *how* a casual model explains an aspect of the real world, or how it allows us to investigate the likely effects of real-world factors on real-world phenomena.

Gibbard and Varian recognize – indeed, they welcome – the fact that casual models are unrealistic; but their defence of this lack of realism is itself rather casual:

> When economic models are used in this way to explain casually observable features of the world, it is important that one be able to grasp the explanation. Simplicity, then, will be a highly desirable feature of such models. Complications to get as close as possible a fit to reality will be undesirable if they make the model less possible to grasp. Such complications may, moreover, be unnecessary, since the aspects of the world the model is used to explain are not precisely measured. (Gibbard and Hal Varian 1978: 672)

The suggestion here seems to be that the purpose of a model is to communicate an idea to an audience; simplicity is a virtue because it makes communication easier. But this puts the cart before the horse. What has to be communicated is not just an idea: it is a claim about how things really are, along with reasons for accepting that claim as true. Simplicity in communication has a point only if there is something to be communicated. While granting that Akerlof's and Schelling's models are easy to grasp, we may still ask what exactly we have grasped. How do these models come to be explanations? And explanations of what?

One possible answer is given by Deirdre McCloskey (1983: 502–507), who argues that models are metaphors. According to McCloskey, the modeller's claim is simply that the real world is like the model in some significant respect (p. 502). In evaluating a model, we should ask the same questions as we would when evaluating a metaphor: 'Is it illuminating, is it satisfying, is it apt?' (p. 506). The claim 'models are metaphors' must, I think, be understood as a metaphor in itself. As a metaphor, it is certainly satisfying and apt; but, in relation to our examination of Akerlof's and Schelling's models, just how illuminating is it?

Clearly, Akerlof and Schelling are claiming that the real world is like their models in some significant respects. What is at issue is what exactly these claims amount to, and how (if at all) they can be justified. Translating into McCloskey's language, what is at issue is how illuminating and how apt Akerlof's and Schelling's metaphors are. But this translation of the question does not take us any nearer to an answer.

Gibbard and Varian (1978) come closer to giving an answer to this question (at this stage, I do not say the right answer) when they suggest that models are *caricatures.* The concept of caricature is tighter than that of metaphor, since the ingredients of a caricature must be taken from the corresponding reality. (Compare cartoons – John Bull, the fat, beef-eating yeoman farmer, was originally a caricature of a characteristic Englishman. Although no longer a valid caricature, he is still recognizable as a symbol of, or metaphor for, Englishness.) According to Gibbard and Varian, the assumptions of a model may be chosen 'not to approximate reality, but to exaggerate or isolate some feature of reality' (p. 673). The aim is 'to distort reality in a way that illuminates certain aspects of that reality' (p. 676).

The idea that models are caricatures suggests that models may be able to explain the real world because their assumptions describe certain features of that world, albeit in isolated or exaggerated form. Gibbard and Varian do not pursue this idea very far, but it is taken up in different ways by Hausman (1992: 123–151) and by Uskali Mäki (1992, 1994), whose work will now be discussed.

7. Economics as an Inexact Deductive Science, and the Method of Isolation

I have suggested that Akerlof and Schelling are each pointing to some tendency in the real world, which each claims to explain by means of a model. One way of trying to make sense of the idea of 'tendencies' is by means of what Hausman calls 'implicit *ceteris paribus* clauses'. The underlying idea is that the phenomena of the real world are the product of the interaction of many different causal factors. A tendency (some writers prefer the term 'capacity') is to be understood as the workings of some small subset of these factors.

In order to describe a tendency, we must somehow isolate the relevant subset of factors from the rest. Thus, the description is expressed in counterfactual terms, such as 'in the absence of all other causal factors, L' or 'if all other causal factors are held constant, L' where L is some law-like proposition about the world. Hausman argues that in economics, *ceteris paribus* clauses are usually both implicit and vague. He uses the term inexact generalization for generalizations that are qualified by implicit *ceteris paribus* clauses.

Hausman argues that economics arrives at its generalizations by what he calls the inexact deductive method. He summarizes this method as the following four-step schema:

1. *Formulate* credible (*ceteris paribus*) and pragmatically convenient generalizations concerning the operation of relevant causal variables;
2. *Deduce* from these generalizations, and statements of initial conditions, simplifications, etc., predictions concerning relevant phenomena;
3. *Test* the predictions;
4. If the predictions are correct, then regard the whole amalgam as confirmed. If the predictions are not correct, then *compare* alternative accounts of the failure on the basis of explanatory success, empirical progress, and pragmatic usefulness (p. 222).

For Hausman, this schema is 'both justifiable and consistent with existing theoretical practice in economics, insofar as that practice aims to appraise theories empirically' (p. 221).[12] By following this schema, economists can arrive at inexact generalizations about the world, which they are entitled to regard as confirmed. The schema is an adaptation of John Stuart Mill's (1843, Book 6, chs 1–4) account of the 'logic of the moral sciences'. (The most significant amendment is that, in Hausman's schema, the premises from which deductions are made are merely 'credible generalizations' which may be called into question if the predictions derived from them prove false. In contrast, Mill seems to have thought that the inexact predictions of economics could be deduced from proven 'laws of mind'.)

Mäki's account of how economic theories explain reality has many similarities with Hausman's. Like Hausman, Mäki argues that theoretical assumptions should be read as claims about what is true in the real world. But where Hausman talks of *inexact* propositions, Mäki talks of *isolations.* Economics, according to Mäki, uses 'the method of isolation, whereby a set of elements is theoretically removed from the influence of other elements in a given situation' (1992: 318). On this account, a theory represents just some of the factors which are at work in the real world; the potential influence of other factors is 'sealed off' (p. 321). Such sealing-off makes a theory unrealistic; but the theory may still claim to describe an aspect of reality.

As Mäki (p. 325) notices, there is a parallel between his concept of theoretical isolation and the idea of *experimental* isolation. Laboratory experiments investigate particular elements of the world by isolating them; the mechanisms by which other elements are sealed off are experimental controls. The laboratory environment is thereby made unrealistic, in the sense that it is 'cleaner' than the world outside; but this unrealisticness is an essential feature of the experimental method. On this analogy, models are *thought experiments.*[13]

But if a thought experiment is to tell us anything about the real world (rather than merely about the structure of our own thoughts), our reasoning must in some way replicate the workings of the world. For example, think how a structural engineer might use a theoretical model to test the strength of a new design. This kind of modelling is possible in engineering because the theory which describes the general properties of the relevant class of structures is already known, even though its implications for the new structure are not. Provided the predictions of the general theory are true, the engineer's thought experiment replicates a physical experiment that could have been carried out.

On this interpretation, then, a model explains reality by virtue of the truth of the assumptions that it makes about the causal factors it has isolated. The isolations themselves may be unrealistic; in a literal sense, the assumptions which represent these isolations may be (and typically are) false. But the assumptions which represent the workings of the isolated causal factors need to be true. So, I suggest, the implications of the method of isolation for theoretical modelling are broadly similar to the first two steps of Hausman's schema. That is, the modeller has to formulate credible generalizations concerning the operation of the factors that have been isolated, and then use deductive reasoning to work out what effects these factors will have in particular controlled environments.

So is this what Akerlof and Schelling are doing? Even though neither author explicitly proposes a testable hypothesis, we might perhaps interpret them as implicitly proposing *ceteris paribus* hypotheses. (Later, I shall suggest what these hypotheses might be.) But if Akerlof's and Schelling's models are to be understood as instances of the inexact deductive method, each model must be interpreted as the deductive machinery which generates the relevant hypothesis. For such an interpretation to be possible, we must be able to identify the simplifying assumptions of the model with the *ceteris paribus* or non-interference clauses of the hypothesis. That is, if the hypothesis takes the form 'X is the case, provided there is no interference of types $i_1, \dots, i_n$', then the model must deduce X from the conjunction of two sets of assumptions. The first set contains 'credible and pragmatically convenient generalizations' – preferably ones which have been used successfully in previous applications of the inexact deductive method. The second set of assumptions – which Mäki would call 'isolations' – postulate the non-existence of $i_1, \dots, i_n$.

Take Akerlof's model. Can its assumptions be understood in this way? Some certainly can. For example, Akerlof implicity assumes that each trader maximizes expected utility. Correctly or incorrectly, most economists regard expected utility maximization as a well-grounded generalization about human behaviour; there are (it is thought) occasional exceptions, but these can safely be handled by implicit non-interference clauses. Similarly, Akerlof assumes that if an equilibrium price exists in a market, that price will come about, and the market will clear. This, too, is a generalization that most economists regard as well-grounded. There is a standing presumption in economics that, if an empirical statement is deduced from standard assumptions such as expected utility maximization and market-clearing, then that statement is reliable: the theorist does not have to justify those assumptions anew in every publication.

As an example of the other type of assumption, notice that Akerlof's model excludes all of the 'countervailing institutions' which he discusses in his section IV. Presumably, if Akerlof is proposing an empirical hypothesis, it must be something like the following: 'If sellers know more than buyers about the quality of a good, and if there are no countervailing institutions, then the average quality of those goods that are traded is lower than that of goods in general.' The absence of countervailing institutions is a non-interference clause in the hypothesis, and therefore also a legitimate property of the model from which the hypothesis is deduced.

The difficulty for a Hausman-like or Mäki-like interpretation is that Akerlof's and Schelling's models both include many assumptions which neither are well-founded generalizations nor correspond with *ceteris paribus* or non-interference clauses in the empirical hypothesis that the modeller is advancing. Akerlof assumes that there are only two types of trader, that all traders are risk-neutral, that all cars are alike except for a one-dimensional index of quality, and so on. Schelling assumes that all individuals are identical except for colour, that they live in the squares of a rectangular grid, and so on again. These are certainly not well-founded empirical generalizations. So can they be read as *ceteris paribus* clauses?

If we are to interpret these assumptions as ceteris paribus clauses, there must be corresponding restrictive clauses in the hypotheses that are deduced from the models. That is, we must interpret Akerlof and Schelling as proposing counterfactual empirical hypotheses about what would be observed, were those assumptions true. But if we pursue the logic of this approach, we end up removing almost all empirical content from the implications of the models – and thereby defeating the supposed objective of the inexact deductive method. Take the case of Schelling's model. Suppose we read Schelling as claiming that *if* people lived in checkerboard cities, and *if* people came in just two colours, and *if* each person was content provided that at least a third of his neighbours were the same colour as him, and *if* . . . , and *if* . . . (going on to list all the properties of the model), *then* cities would be racially segregated. That is not an empirical claim at all: it is a theorem.

Perhaps the best way to fit Akerlof's and Schelling's models into Hausman's schema is to interpret their troublesome assumptions as the 'simplifications etc.' referred to in step 2 of that schema. But this just shunts the problem on, since we may then ask why it is legitimate to introduce such simplifications into a deductive argument. The conclusions of a deductive argument cannot be any stronger than its premises. Thus, any hypothesis that is generated by a deductive method must have implicit qualifying clauses corresponding with the assumptions that are used as premises. And this does not seem to be true of Akerlof's and Schelling's hypotheses.

To understand what Akerlof and Schelling are doing, we have to realize that results that they derive deductively within their models are not the same as the hypotheses that they want us to entertain. Consider exactly what Akerlof and Schelling are able to show by means of their models. Akerlof shows us that under certain specific conditions (there are just two types of trader, all cars are identical except for quality, sellers' valuations of cars of given quality are two-thirds those of buyers, etc.), no trade takes place. Among these conditions is a particular assumption about asymmetric information: sellers know the quality of their cars, but buyers don't. Akerlof also shows that if the only change that is made to this set of conditions is to assume symmetric information instead of asymmetric, then trade does take place. Thus, Akerlof has proved a *ceteris paribus* result, but only for a particular array of other conditions. This result might be roughly translated as the following statement: If all other variables are held constant at the particular values assumed in the model, then an increase in the degree of asymmetry of information reduces the volume of trade.

What about Schelling? Schelling shows – or, strictly speaking, he invites us to show ourselves – that under certain specific conditions (people come in just two colours, each person is located on a checkerboard, etc.) individuals' independent choices of location generate segregated neighbourhoods. Among these conditions is a particular assumption about individuals' preferences concerning the colour composition of their neighbourhoods: people prefer not to live where more than some proportion p of their neighbours are of the other colour. Schelling invites us to try out different values of p. We find that segregated neighbourhoods eventually evolve, whatever value of p we use, provided it is less than 1. If $p = 1$, that is, if people are completely indifferent about the colours of their neighbours, then segregated neighbourhoods will not evolve. (Schelling does not spell out this latter result, but a moment's thought about the model is enough to derive it.) Thus, we have established a *ceteris paribus* result analogous with Akerlof's: we have discovered the effects of changes in the value of p, when all other variables are held constant at the particular values specified by the model.

To put this more abstractly, let x be some variable whose value we are trying to explain, and let $(v_1, \ldots, v_n)$ be an array of variables which might have some influence on x. What Akerlof and Schelling each succeed in establishing by deductive reasoning is the truth of a proposition of the form: If the values of $v_2, \ldots, v_n$ are held constant at the specific values $v_2^*, \ldots, v_n^*$, then the relationship between v_1 and x is. . . . The values $v_2^*, \ldots, v_n^*$ are those built into the relevant model. Taken at face value, this proposition tells us nothing about the relationship between v_1 and x in the actual world. It tells us only about that relationship in a counterfactual world.

But Akerlof and Schelling want us to conclude that certain much more general propositions are, if not definitely true, at least credible. When Akerlof talks about the 'lemons principle', he has in mind some broad generalization, perhaps something like the following: For all markets, if all other features are held constant, an increase in the degree of asymmetry of information reduces the volume of trade. Similarly, what Schelling has in mind is some generalization like the following: For all multi-ethnic cities, if people prefer not to live in neighbourhoods where the vast majority of their neighbours are of another ethnic group, strongly segregated neighbourhoods will evolve. In my more abstract notation, the generalizations that Akerlof and Schelling have in mind have the form: If the values of $v_2, \ldots, v_n$ are held constant at any given value, then the relationship between v_1 and x is. . . .

If these generalizations are to be interpreted as hypotheses, the models are supposed to give us reasons for thinking that they are true. If the generalizations are to be interpreted as observed regularities, the models are supposed to explain why they are true. But deductive reasoning cannot fill the gap between the specific propositions that can be shown to be true in the model world (that is, propositions that are true if $v_2, \ldots, v_n$ are held constant at the values $v_2^* \ldots, v_n^*$) and the general propositions that we are being invited to entertain (that is, those that are true if $v_2, \ldots, v_n$ are held constant at any values). Somehow, a transition has to be made from a particular hypothesis, which has been shown to be true in the model world, to a general hypothesis, which we can expect to be true in the real world too.

8. Inductive Inference

So how can this transition be made? As before, let R stand for a regularity (bad products driving out good, persistent racial segregation with moving geographical boundaries) which may or may not occur in the real world. Let F stand for a set of causal factors (sellers being better-informed than buyers, a common preference not to be heavily outnumbered by neighbours not of one's own type) which may or may not operate in the real world. Akerlof and Schelling seem to be reasoning something like this:

Schema 1: Explanation
E1 – in the model world, R is caused by F.
E2 – F operates in the real world.
E3 – R occurs in the real world.
Therefore, there is reason to believe:
E4 – in the real world, R is caused by F.

Alternatively, if we read Akerlof and Schelling as implicitly proposing empirical hypotheses, we might represent their reasoning as:

Schema 2: Prediction
P1 – in the model world, R is caused by F.
P2 – F operates in the real world.
Therefore, there is reason to believe:
P3 – R occurs in the real world.

A third possible reading of Akerlof and Schelling involves abductive reasoning (inferring causes from effects):[14]

Schema 3: Abduction
A1 – in the model world, R is caused by F.
A2 – R occurs in the real world.
Therefore, there is reason to believe:
A3 – F operates in the real world.

In each of these three reasoning schemata, the 'therefore' requires an inductive leap. By 'induction' I mean any mode of reasoning which takes us from specific propositions to more general ones (compare the similar definition given by Mill [1843, Book 3, ch. 1, p. 186]). Here, the specific proposition is that R is caused by F in the case of the model. In order to justify each of the 'therefores', we must be justified in inferring that R is caused by F more generally. *If* there is a general causal link running from F to R, then when we observe F and R together in some particular case (that is, the case of the real world), we have some reason to think that the particular R is caused by the particular F (explanation). Similarly, when we observe F in a particular case, we have some reason to expect to find R too (prediction). And when we observe R in a particular case, we have some reason to expect to find F too (abduction). It seems, then, that Akerlof's and Schelling's method is not purely deductive: it depends on induction as well as on deduction. But how might these inductions be justified?

9. Justifying Induction: Separability

One possible answer is to appeal to a very general hypothesis about causation, which (to my knowledge) was first invoked by Mill (1843, Book 3, ch. 6, pp. 242–247). Mill defines phenomena as mechanical if the overall effect of all causal factors can be represented as an addition of those separate factors, on the analogy of the vector addition of forces in Newtonian physics. Given this hypothesis of the composition of causes, we are entitled to move from the *ceteris paribus* propositions which have been shown to be true in a model to more general *ceteris paribus* propositions which apply to the real world too.[15] Using the notation introduced in section 6, this immediately closes the gap between a proposition which is true if certain variables $v_2, \ldots, v_n$ are held constant at certain specific values $v_2^*, \ldots, v_n^*$ and a proposition which is true if $v_2, \ldots, v_n$ are held constant at any values: if the proposition is true in the first case, then (if the hypothesis about the composition of causes is true) it is true in the second case too. But what entitles us to use that hypothesis itself?

In some cases, it may be legitimate to treat that hypothesis as a proven scientific law – as in the paradigm case of the composition of forces in physics. Mill seems to have taken it to be an *a priori* truth that 'In social phenomena the Composition of Causes is the universal law' (1843, Book 6, ch. 7, p. 573). However, the argument Mill gives in support of this claim is quite inadequate. He simply asserts that 'Human beings in society have no properties but those which are derived from, and may be resolved into, the laws of the nature of individual man'. But even if we grant this assertion, all we have established is that social facts are separable into facts about individuals. We have not established the separability of *causal factors*. Thus, for example, the fact that society is an aggregate of individuals does not allow us to deduce that if an increase in the price of some good in one set of circumstances causes a decrease in consumption, then the same cause will produce the same effect in other circumstances.

Hausman (1992: 138) offers a defence for Mill's method in economics. He claims that Mill's supposition that economic phenomena are mechanical is 'implicit in most applications of economic models', and then says: 'Its only justification is success'. In other words, this supposition is an inductive inference from the general experience of economic modelling.

But this argument seems to beg the question. For the sake of the argument, let us grant that economic modelling has often been successful – successful, that is, in relation to Hausman's criterion of generating correct predictions about the real world. Even so, the explanation of its success may be that economists are careful not to rely on models unless they have some independent grounds for believing that the *particular* phenomena they are trying to explain are mechanical – or, more generally, unless they have some independent grounds for making particular inductive inferences from the world of the model to the real world. Given the *prima facie* implausibility of the assumption that all economic phenomena are mechanical, it would be surprising to find that this assumption was the main foundation for inductive inferences from theoretical models. We should look for other foundations.

10. Justifying Induction: Robustness

One way in which inductions might be justified is by showing that the results derived from a model are *robust* to changes in the specification of that model. Gibbard and Varian (1978: 675) appeal to the robustness criterion when they suggest that, in order for caricature-like models to help us to understand reality, ‘the conclusions [should be] robust under changes in the caricature’. Hausman (1992: 149) makes a somewhat similar appeal when he considers the conditions under which it is legitimate to use simplifications – that is, propositions that are not true of the real world – in the second stage of his schema of the inexact deductive method. He proposes a set of conditions which he glosses as ‘reasonable criteria for judging whether the falsity in simplifications is irrelevant to the conclusions one derives with their help’.

One significant implication of this approach is that simplifications need not be isolations. Take Schelling’s checkerboard city. The simplicity of the checkerboard city lies in the way that its pattern repeats itself: if we ignore the edges of the board, every location is identical with every other. (More showy theorists than Schelling would probably draw the checkerboard on a torus, so that it had no edges at all; this would give us a city located on a doughnut-shaped planet.) This property of ‘repeatingness’ makes the analysis of the model much easier than it otherwise would be. But it does not seem right to say that the checkerboard *isolates* some aspect of real cities by sealing off various other factors which operate in reality: just what do we have to seal off to make a real city – say, Norwich – become like a checkerboard? Notice that, in order to arrive at the checkerboard plan, it is not enough just to suppose that all locations are identical with one another (that is, to use a ‘generic’ concept of location): we need to use a *particular form* of generic location. So, I suggest, it is more natural to say that the checkerboard plan is something that Schelling has *constructed* for himself. If we think that Schelling’s results are sufficiently robust to changes in the checkerboard assumption, that assumption may be justified, even though it is not an isolation.[16]

Robustness arguments work by giving reasons for believing that a result that has been derived in one specific model would also be derived from a wide class of models, or from some very general model which included the original model as a special case. Economic theorists tend to like general models, and much effort is put into generalizing results. By experience, theorists pick up a feel for the kinds of result that can be generalized and the kinds that cannot be. The main way of making this distinction, I think, is to examine the links between the assumptions of a model and its results, and to try to find out which assumptions are (as theorists say) ‘doing the work’. If a model has already been presented in a somewhat general way, it is often useful to strip it down to its simplest form, and then to see which assumptions are most closely associated with the derivation of the relevant result.[17]

In both Akerlof's and Schelling's models, there are good reasons to think that most of the simplifying assumptions are orthogonal to the dimension on which the model 'works': these are simplifying assumptions which could be changed or generalized without affecting the qualitative results. In many cases, Akerlof argues exactly this. Recall, for example, his discussion of risk neutrality. Akerlof could have assumed risk aversion instead, which would have made the model much less easy to work with; but there does not seem to be any way in which the major qualitative conclusions are being driven by the assumption of risk neutrality. Similarly, in the case of Schelling's model, the checkerboard layout seems to have nothing particular to do with the tendency for segregation. Schelling is confident enough to invite the reader to try different shapes of boards, and might easily have suggested different tessellations (such as triangles or hexagons).

Notice how this mode of reasoning remains in the world of models – which may help to explain why theorists feel comfortable with it. It makes inductive inferences from one or a small number of models to *models* in general. For example: having experimented with Schelling's checkerboard model with various parameter values, I have found that the regularity described by Schelling persistently occurs. Having read Schelling and having thought about these results, I think I have some feel for why this regularity occurs; but I cannot give any proof that it *must* occur (or even that it must occur with high probability). My confidence that I would find similar results were I to use different parameter values is an inductive inference. I also feel confident (although not quite as confident as in the previous case) that I would find similar results if I used triangles or hexagons instead of squares. This is an inductive inference too.

Obviously, however, it cannot be enough to stay in the world of models. If the theorist is to make claims about the real world, there has to be some link between those two worlds. For example, it is not enough to be convinced that what Schelling has shown us to be true of checkerboard cities is also true of other modelcities: we have to be convinced that it is true of real cities. We have to think something like the following: If what Schelling has shown us is true of checkerboard cities, then it will probably tend to be true of cities in general. What makes that inductive inference credible?

11. Justifying Induction: Credible Worlds

Inductive reasoning works by finding some regularity R in some specific collection of observations $x_1, \ldots, x_n$, and then inferring that the same regularity will probably be found throughout a general set of phenomena S, which contains not only $x_1, \ldots, x_n$ but also other elements which have not yet been observed. For example, $x_1, \ldots, x_n$ might be the n different versions of Schelling's checkerboard city that I have so far experimented with, R might be the emergence of segregation in model cities, and S might be the set of all checkerboard cities. Having found R in the n particular cities, I infer that this is a property of checkerboard cities in general.

Unavoidably, inductive reasoning depends on prior concepts of similarity: we have to be able to interpret S as the definition of some *relevant* or *salient* respect in which $x_1, \ldots, x_n$ are similar. Many of the philosophical puzzles surrounding induction stem from the difficulty of justifying any criterion of similarity.[18] Obviously, I am not going to solve these deep puzzles towards the end of a paper about models in economics.[19] For my purposes, what is important is this: if we are to make inductive inferences from the world of a model to the real world, we must recognize some significant similarity between those two worlds.

If we interpret Akerlof and Schelling as using schema 1 or schema 2 (see section 7), it might be said that this similarity is simply the set of causal factors F: what the two worlds have in common is that those factors are present in both. To put this another way, the real world is equivalent to an immensely complicated model: it is the limiting case of the process of replacing the simplifying assumptions of the original model with increasingly realistic specifications. If (as I argued in section 10) we can legitimately make inductive inferences from a simple model to slightly more complex variants, then we must also have *some* warrant for making inferences to much more complex variants, and hence also to the real world. Nevertheless, the enormous difference in complexity between the real world and any model we can hope to analyse – and hence the apparent lack of similarity between the two – suggests that we ought to be very cautious about making inferences from the latter to the former.

So what might increase our confidence in such inferences? I want to suggest that we can have more confidence in them, the greater the extent to which we can understand the relevant model as a description of how the world *could be.*

Let me explain. Inductive inferences are most commonly used to take us from one part of the real world to another. For example, suppose we observe racial segregation in the housing markets of Baltimore, Philadelphia, New York, Detroit, Toledo, Buffalo and Pittsburgh. Then we might make the inductive inference that segregation is a characteristic of large industrial cities of the north-eastern USA, and so form the expectation that there will be segregation in say, Cleveland. Presumably, the thought behind this inference is that the forces at work in the Cleveland housing market, whatever these may be, are likely to be broadly similar to those at work in other large industrial cities in north east USA. Thus, a property that is true for those cities in general is likely to be true for Cleveland in particular. One way of describing this inference is to say that each of the housing markets of Baltimore, Philadelphia, New York, etc. constitutes a *model* of the forces at work in large industrial north-eastern US cities. These, of course, are natural models, as contrasted with *theoretical* models created in the minds of social scientists. But if we can make inductive inferences from natural models, why not from theoretical ones? Is the geography of Cleveland any more like the geography of Baltimore or Philadelphia than it is like the geography of Schelling's checkerboard city?[20]

What Schelling has done is to construct a set of *imaginary* cities, whose workings we can easily understand. In these cities, racial segregation evolves only if people have preferences about the racial mix of their neighbours, but strong segregation evolves even if those preferences are quite mild. In these imaginary cities, we also find that the spatial boundaries between the races tend to move over time, while segregation is preserved. We are invited to make the inductive inference that similar causal processes apply in real multi-ethnic cities. We now look at such cities. Here too we find strong spatial segregation between ethnic groups, and here too we find that the boundaries between groups move over time. Since the same effects are found in both real and imaginary cities, it is at least credible to suppose that the same causes are responsible. Thus, we have been given some reason to think that segregation in real cities is caused by preferences for segregation, and that the extent of segregation is invariant to changes in the strength of those preferences.

Compare Akerlof. Akerlof has constructed two variants of an imaginary used-car market. In one variant, buyers and sellers have the same imperfect information about the quality of cars, and trade takes place quite normally. In the other variant, sellers know more than buyers, and no trade takes place at all. When we think about how these markets work, it becomes credible to suppose that many variant imaginary markets can be constructed, and that these share the common feature that, *ceteris paribus*, the volume of trade falls as information becomes less symmetric. We are invited to make the inductive inference that similar causal processes apply in real markets, with similar effects. Thus in real markets too, *ceteris paribus*, the volume of trade is positively related to the symmetry of information.

We gain confidence in such inductive inferences, I suggest, by being able to see the relevant models as instances of some category, some of whose instances actually exist in the real world. Thus, we see Schelling's checkerboard cities as *possible cities*, alongside real cities like New York and Philadelphia. We see Akerlof's used-car market as a *possible market*, alongside real markets such as the real market for used cars in a particular city, or the market for a particular type of insurance. We recognize the significance of the similarity between model cities and real cities, or between model markets and real markets, by accepting that the model world *could be* real – that it describes a state of affairs that is *credible*, given what we know (or think we know) about the general laws governing events in the real world. On this view, the model is not so much an abstraction from reality as a parallel reality. The model world is not constructed by starting with the real world and stripping out complicating factors: although the model world is simpler than the real world, the one is not a *simplification* of the other.

Credibility in models is, I think, rather like credibility in 'realistic' novels. In a realistic novel, the characters and locations are imaginary, but the author has to convince us that they are credible – that there could be people and places like those in the novel. As events occur in the novel, we should have the sense that these are natural outcomes of the way the characters think and behave, and of the way the world works. We judge the author to have failed if we find a person acting out of character, or if we find an anachronism in a historical novel: these are things that *couldn't* have happened. But we do not demand that the events of the novel did happen, or even that they are simplified representations of what really happened. (Simplification and isolation are allowed, of course; we do not expect to be told everything that the characters do or think. But what is being simplified is not the world of actual events, but the world imagined by the author.) We can praise a novel for being 'true to life' while accepting that every event within it is fictional, as when we recognize aspects of its characters as typical of people we know. When a novel has this form of truth, we can even use it to explore 'What would happen if . . . ?' questions, in something like the same way that economists can use models. By following the characters' reactions to events that we have not ourselves experienced, we may gain insights into how we would react in similar circumstances.[21]

But the reader will expect more than analogy. The obvious question that I have to answer is: What constitutes credibility in economic models? I cannot give anything remotely like a complete answer; the best I can offer are a few criteria that have guided me in my own work as a modeller, and which are exemplified in the economic models that I most admire.

For me, one important dimension of credibility is *coherence*. Everyone recognizes that a theoretical model has to be *logically* coherent, but I mean something more than this. The assumptions of a good model cohere in the broader sense that they fit naturally together. For example, some economic models assume that agents are well-informed and highly rational, while others assume that agents are poorly-informed and follow rough rules of thumb. Which type of model is more useful in explaining particular phenomena is a matter of judgement. But a model which uses an apparently arbitrary mix of the two kinds of assumption – assuming hyper-rationality in one context and bounded rationality in another – has the same kind of fault as a novel in which someone acts out of character. If a model lacks coherence, its results cannot be seen to follow naturally from a clear conception of how the world might be; this prompts the suspicion that the assumptions have been cobbled together to generate predetermined results. *Ad hoc* models of this kind may be commonplace in economics journals, but if they are, that does not justify them.

For a model to have credibility, it is not enough that its assumptions cohere with one another; they must also cohere with what is known about causal processes in the real world. Thus, Akerlof's assumption that prices tend to their market-clearing levels is justified by evidence from a wide range of 'natural' and laboratory markets. Schelling's assumption that many people have at least mildly segregationist preferences is justified by psychological and sociological evidence, and coheres with common intuition and experience. However, it is not necessary that the assumptions of the model correspond with – or even with a simplification of – any *particular* real-world situation. Thus, we should not object to Akerlof's assumption that traders' utility functions are additively separable in money and the quality of cars, or his assumption that cars are worth exactly 50 per cent more to traders of one type than they are to traders of another. These are *restrictive* assumptions, but they seem adequately *representative* of people who trade cars in the real world. In the same way, the author of a novel might choose to call her principal character Frank, make him 48 years old, and fix his home town as Ipswich. If the logic of the novel requires only that the principal character is middle-aged, male and English, there is a sense in which this specification is highly restrictive; but the character has to have *some* name, *some* age, and *some* home town, and this particular specification is adequately representative of middle-aged English men (whereas, say, naming the character Duck Bill Platypus is not).

Akerlof in particular puts a lot of effort into making his model credible in the sense I have tried to describe. The world of his model is much more uniform and regular than the real world, but Akerlof clearly wants us to think that there *could* be a used-car market which was like his model. The 'cars' and 'traders' of his model are not just primitives in a formal deductive system. They are, I suggest, cars which are *like* real cars, and traders which are *like* real traders, inhabiting a world which Akerlof has imagined, but which is sufficiently close to the real world that we can imagine its being real. Recall the sentence in which Akerlof seems to slip between talking about the real used-car market and talking about his model: the fact that such slippage is possible may be an indication that Akerlof has come to think of his model as if it were real.

At first sight, Schelling seems rather less concerned to make us believe in his model world as a possible reality. Instead of following Akerlof's strategy of basing his model on one typical case, Schelling almost always refers to the two types of actor in his model as 'dimes' and 'pennies'. But this is perhaps dictated by Schelling's strategy of asking the reader to perform the actions in the model: he has to say 'now move that dime' rather than 'that dime now moves'. Possibly, too, it reflects an embarrassment about dealing directly with the issue of racial prejudice. But when Schelling describes the laws of motion of these coins, it is clear that we are expected to think of them as people. For example, one of his suggestions is that 'we can postulate that every dime wants at least half its neighbours to be dimes, every penny wants a third of its neighbours to be pennies, and any dime or penny whose immediate neighbourhood does not meet these conditions gets up and moves' (pp. 147–148). Or again, officially referring to a dime or penny in a world of dimes and pennies: 'He is content or discontent with his neighbourhood according to the colours of the occupants of those eight surrounding squares . . . ' (p. 148). Even allowing for the fact that the use of 'he' and 'colour' rather than 'it' and 'type of coin' are probably slips, it is surely obvious that Schelling wants us to think of the dimes and pennies as people of two groups who have some embarrassment about being together. Similarly, we are expected to think of the checkerboard as a city (or some other social space, such as a dining room). Further, we are encouraged to think of these people's attitudes to one another as credible and understandable – even forgivable (recall the passage about mixed tables in the cafeteria, which precedes the checkerboard model). What Schelling has constructed is a model city, inhabited by people who are *like* real people.

12. Conclusion

I have referred several times to a puzzling common feature of the two papers. Both authors seem to want to make empirical claims about properties of the real world, and to want to argue that these claims are supported by their models. But on closer inspection of the texts, it is difficult to find any explicit connection being made between the models and the real world. Although both authors discuss real-world phenomena, neither seems prepared to endorse any specific inference from his model, still less to propose an explicit hypothesis which could be tested.

I suggest that the explanation of this puzzle is that Akerlof and Schelling are engaged in a kind of theorizing the usefulness of which depends on inductive inferences from the world of models to the real world. Everyone makes inductive inferences, but no one has really succeeded in justifying them. Thus, it should not be surprising if economists leave gaps in their explicit reasoning at those places where inductive inferences are required, and rely on their readers using their own intuitions to cross those gaps. Nor should it be surprising if economists use rhetorical devices which tend to hide these gaps from view.

Nevertheless, the gap between model and real world has to be bridged. If a model is genuinely to tell us something, however limited, about the real world, it cannot be *just* a description of a self-contained imaginary world. And yet theoretical models in economics often *are* descriptions of self-contained and imaginary worlds. These worlds have not been formed merely by abstracting key features from the real world; in important respects, they have been *constructed* by their authors.

The suggestion of this paper is that the gap between model world and real world can be filled by inductive inference. On this account, models are not internally consistent sets of uninterpreted theorems; but neither are they simplified or abstracted or exaggerated descriptions of the real world. They describe credible counterfactual worlds. This credibility gives us some warrant for making inductive inferences from model to real world.

Acknowledgements

A previous version of this paper was prepared for the conference *Fact or Fiction? Perspectives on Realism and Economics* at the Erasmus Institute for Philosophy and Economics, Rotterdam, in November 1997. The paper has been much improved as a result of the discussion at that conference. I particularly thank Nancy Cartwright, Stephan Hartmann, Daniel Hausman, Maarten Janssen, Uskali Mäki, Mary Morgan and Chris Starmer for advice. I did most of the work on the paper while visiting the Centre for Applied Ethics at the University of British Columbia, for whose hospitality I am grateful. Subsequent work was supported by the Leverhulme Trust.

Notes

1. But it was not immediately recognized as a major contribution: it was turned down three times before being accepted for publication. Mark Blaug (1997) uses this fact to suggest that Akerlof's paper is the exception which proves the rule – the rule being that modern economics is becoming 'an intellectual game played for its own sake and not for its practical consequences', creating models which are 'scandalously unrepresentative of any recognizable economic system' (pp. 2–4). However, he does not explain why Akerlof is to be acquitted of this charge.
2. An alternative reading is possible. Akerlof never claims outright that the 'pure 'joy' explanation is false, or that his own explanation is correct – only that it is 'different'. So could it be that he doesn't want to make any such claims? In section 3, I consider – and reject – the suggestion that Akerlof is not claiming to explain any features of the real world.
3. Akerlof deals with this problem to some degree by sketching a model with four discrete types of car. (This sketch is contained in the passage beginning 'Suppose . . . '.) In the four-types model, there is a market in bad used cars but not in good ones. However, this model is not developed in any detail; it serves as a kind of appetizer for the main model, in which no trade takes place at all.
4. As a result of presenting this paper, I have discovered that Schelling's model is much more widely known and admired than I had imagined. It has not had the obvious influence on economics that Akerlof's paper has, but it clearly appeals to methodologically-inclined economists.
5. In passing, I must record my puzzlement at the two-way classification of 'colours' or 'races' which seems to be a social fact in America, despite the continuity of the actual spectra of skin colour, hair type and other supposed racial markers. The convention, I take it, is that anyone of mixed African and European parentage, whatever that mix, is black unless he or she can 'pass' as pure European.
6. When I have presented this paper, I have been surprised at how many economists are inclined towards this interpretation.
7. Arrow (1951: 4–5) hints at this interpretation when, as part of the introduction to his presentation of the theorem, he says that welfare economists need to check that the value judgements they invoke are mutually compatible. He goes on: 'Bergson considers it possible to establish an ordering of social states which is based on the indifference maps of individuals, and Samuelson has agreed'. Arrow's form of social choice theory investigates whether this is indeed possible.
8. This interpretation of Akerlof's model was suggested to me by Daniel Hausman. Hausman also suggested the 'counter-example' interpretation of Schelling's model, discussed in the next paragraph.
9. Here I am using 'story' in the sense which McCloskey (1983: 505) correctly identifies as standard usage among economic theorists: 'an extended example of the economic reasoning underlying the mathematics [of a theory], often a simplified version of the situation in the real world that the mathematics is meant to characterize'. Gibbard and Varian (1978) use 'story' in a similar way (see section 6). Morgan (1997) has a quite different concept of a story. For Morgan, models are inert mechanisms which need to be 'cranked' by some external event

in order to set them in motion; a story is a description of that event and of how its impact is transmitted through the model. Morgan's approach conflates two distinctions – static/dynamic and model/story – which I prefer to keep separate.

10. Early astronomy provides a classic example of the conflict between instrumentalism and realism. The only available observations were of the movements of points and areas of light across the sky. Highly accurate predictions of these movements could be made by using theories based on apparently fantastic and (at the time) completely unverifiable assumptions about how the workings of the universe might look, viewed from outside. With hindsight, we know that some of these fantastic assumptions proved to be true (which supports realism), while others proved false (which supports instrumentalism).
11. The idea that there might be some value in predicting the consumption decisions of individual consumers would perhaps not occur to an economist in the 1950s or 1960s, when the instrumentalist defence of neoclassical theory was most popular. At that time, there were no practicable means to collect or to analyse individual-level data. Developments in retailing and in information technology are now opening up the possibility of making profitable use of predictions about the decisions of individual consumers.
12. Hausman adds the qualification that 'a great deal of theoretical work in economics is concerned with conceptual exploration, not with empirical theorizing' (p. 221). In section 4, I considered and rejected the suggestion that Akerlof's and Schelling's models could be interpreted as conceptual explanation.
13. The parallel between models and experiments is explored in detail by Guala (1999).
14. This interpretation was suggested to me by Maarten Janssen.
15. Cartwright (1998) explores the role of this kind of reasoning in Mill's scientific method.
16. There is an analogy in experimental method. Think of how experimental biologists use fruit flies to test and refine hypotheses about biological evolution. The hypotheses in which the biologists are interested are intended to apply to many species other than fruit flies – sometimes, for example, to humans. Fruit flies are used because they are easy to keep in the laboratory and breed very quickly. But fruit flies are not simplified versions of humans, arrived at by isolating certain key features. Rather, the biologist's claim is that certain fundamental evolutionary mechanisms are common to humans and fruit flies.
17. Akerlof and Schelling are perhaps atypical in that they are satisfied to present simple, imaginative models, leaving it to the technicians of economic theory to produce the generalizations. In contrast, most theorists feel compelled to present their models in the most general form they can. If I am right about the importance of stripping down a model in order to judge how generalizable it is, it is at least arguable that Akerlof's and Schelling's way of presenting models is the more informative.
18. The 'grue' problem discovered by Nelson Goodman (1954) is particularly significant – and intractable.
19. For what it is worth, I am inclined to agree with David Hume's (1740, Book 1, Part 3, pp. 69–179) original diagnosis: that induction is grounded in associations of ideas that the human mind finds natural. If that diagnosis is correct, the

concepts of similarity which underpin inductive reasoning may be capable of being explained in psychological terms, but not of being justified as rational.

20. Notice that one implication of thinking in this way is that regularities within the real world (here, across cities which in many respects are very different from one another) can give us grounds for greater confidence in inductive inferences from a model to the real world. The fact that racial segregation is common to so many different cities suggests that its causes are not to be found in any of those dimensions on which they can be differentiated.
21. I still recall the deep impression made on me as a teenager by Stan Barstow's *A Kind of Loving*. The main character of this classic of northern English realistic fiction is a very ordinary young man who gets his girlfriend pregnant and is then pushed into an unwanted marriage. Reading this book, I gained a vivid sense of the possible consequences for me of actions that I could imagine myself taking.

References

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

Arrow, K. J. (1963) *Social Choice and Individual Values*, 2nd edn, New Haven, CT: Yale University Press. (1st edn 1951.)

Blaug, M. (1997) 'Ugly currents in modern economics', paper presented at conference *Fact or Fiction? Perspectives on Realism and Economics*, Erasmus University, Rotterdam, November 1997, and in Uskali Mäki (ed.) *Fact and Fiction. Foundational Perspectives on Economics and the Economy*, forthcoming.

Cartwright, N. (1998) 'Capacities', forthcoming in *The Handbook of Methodology*, Aldershot: Edward Elgar.

Gibbard, A. and Varian, H. (1978) 'Economic Models', *Journal of Philosophy* 75: 664–677.

Goodman, N. (1954) *Fact, Fiction, and Forecast*, Cambridge, MA: Harvard University Press.

Guala, F. (1999) 'Economics and the Laboratory', Ph.D thesis, London School of Economics and Political Science.

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

Hume, D. (1740) *A Treatise of Human Nature*, page references to 1978 edn, Oxford: Clarendon Press.

McCloskey, D. (1983) 'The Rhetoric of Economies', *Journal of Economic Literature* 21: 481–517.

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

Mäki, U. (1994) 'Isolation, Idealization and Truth in Economics', *Poznań Studies in the Philosophy of the Sciences and the Humanities* 38: 147–168.

Mill, J. S. (1843) *A System of Logic*, page references to 1967 edn, London: Longman.

Morgan, M. S. (1997) 'Models, Stories and the Economic World', paper presentepresented at conference *Fact or Fiction? Perspectives on Realism and Economics*, Erasmus University, Rotterdam, November 1997, and in Uskali Mäki (ed.) *Fact and Fiction. Foundational Perspectives on Economics and the Economy*, forthcoming.

Schelling, T. C. (1978) *Micromotives and Macrobehaviour*, New York: Norton.