Journal of Economic Methodology 8:1, 113–130 2001

The evolutionary turn in game theory

Robert Sugden

1 INTRODUCTION

In a conference bar a few years ago, I was talking with an economist who specializes in applications of game theory.[1] I mentioned that I was interested in the conceptual problems associated with the assumption that players have common knowledge of their rationality. His[2] reaction was surprise that anyone still thought seriously about that issue at all. With something like the embarrassment with which people of my age look at photographs of themselves dressed in the fashions of the early 1970s, he confessed that there had been a time, not so very long ago, when he had been engrossed in what has come to be called the refinement programme, in which questions about the meaning of 'common knowledge' and 'rationality' are central. But, of course, not *now*. Now, did not everyone recognize that this programme had been a silly waste of time, and that all the most interesting work in game theory was based on an evolutionary approach?

I was not surprised by this response: it is a commonplace. But at a deeper level, the fact that so many game theorists now say things like this *is* surprising.

In the last ten or so years, game theory has executed a sharp change of direction, away from models based on assumptions of ideal rationality and towards models that are inspired by evolutionary biology. The assumption of individual rationality has usually been interpreted as the most fundamental assumption of economics; to give it up is surely a huge step to take. That biology should be taken as the natural science on which to model economics is itself a major break with tradition. From the neoclassical revolution onwards, theoretical economics have taken *physics*, with its aim of explaining everything in terms of a few simple mathematical principles, as their inspiration. Biology is a much messier discipline; it deals with evolutionary processes that are path-dependent, subject to historical contingencies, and in many respects inherently unpredictable.

Taken at face value, then, the *evolutionary turn* in game theory is a momentous theoretical revolution, a sledge-hammering of hard cores, a

Journal of Economic Methodology ISSN 1350-178X print/ISSN 1469-9427 online
http://www.tandf.co.uk/journals
DOI: 10.1080/1350178001002328 9

dispatch of sacred cows to the abattoir. It might be seen as one of the most fundamental developments in economics in recent years, pointing the path that the discipline is taking at the turn of the century. But then it is puzzling that the whole process has been so *bloodless*. My impression is that there has been little resistance or attempts at counter-revolution: most game theorists, like the one I met at the conference, have switched to the new approach without reservations, indeed with enthusiasm.

I have been involved in another, contemporaneous and in some way parallel, attempt at a theoretical revolution, and the contrast between the two is striking. *Behavioural economics* challenges conventional rationality assumptions on the basis of experimental evidence, and uses ideas from cognitive psychology to guide the construction of new theories. Here, events are unfolding much more as a methodologist of science would expect. There is a lively and hard-fought debate between opposing camps. There are committed defenders of rationality-based theories, who challenge the validity of the experiments carried out by their opponents; the behavioural economists respond by running further experiments with additional controls. It is perhaps too early to say who is winning the debate, but the investment of pride and the commitment of intellectual capital on both sides is obvious to any observer.[3]

The absence of similar passion about the evolutionary turn might be a sign that the arguments in favour of the new approach were overwhelmingly strong, and were quickly recognized as such. But there is an alternative interpretation: perhaps the evolutionary turn is not a theoretical revolution at all, but merely a swing of fashion. If this latter interpretation is the right one, it raises questions about what the fundamental presuppositions of current economic theory are or, indeed, whether there *are* any. If the assumption of individual rationality can be given up without any disruption to the course of economic theory, what assumptions cannot be given up? But then there is another possibility. Perhaps the assumption has not really been given up at all. Perhaps the evolutionary turn is not the fundamental change that it purports to be. Could it be just a restyling of an old model, like the restylings that car manufacturers go in for when a model has begun to age but is not old enough to be replaced?

2 TWO UNSOLVED PROBLEMS IN CLASSICAL GAME THEORY

One clue that the evolutionary turn is not as fundamental as it appears at first sight is given by the way that game theorists often explain the process that led to it. Here I can rely only on what I have heard game theorists say off the record, in conversations and in oral presentations. I hope that most economists will recognize the truth of these observations from their own experience; but I cannot document them. Game theorists are not given to

expressing their methodological and conceptual presuppositions in their published work.

Classical game theory – the version that was standard up to the late 1980s – is built on the assumptions of perfect rationality and common knowledge. The object of the theory is to propose *solutions* for games. A *game* is defined as a mathematical object. The simplest such object is the *normal-form* game, defined in terms of *players*, *strategies* and *utilities*. An *extensive-form* game has players who make decisions at points in a *game tree*; *information sets* describe the players' states of knowledge when they make their decisions. Some formulations of games allow *moves of nature*, with or without associated objective probabilities. For a given game, a solution is a combination of strategies, one for each player. A *solution concept* is a rule which applies to all games in some general class and which, for each such game, picks out one or more combinations of strategies as *the* solution or solutions.

The criteria that distinguish acceptable solution concepts from unacceptable ones are not clearly defined; trying to develop such criteria is part of the enterprise of classical game theory. But there is general agreement that, in order to be acceptable, a solution concept should be compatible with the assumption that players are rational, and that their rationality is a matter of common knowledge among them. There is also a background presumption that, if it is to be useful, a solution concept should, in a typical game, reject a substantial proportion of possible combinations of strategies. The Holy Grail is a solution concept which, for every game, picks out one and only one combination of strategies as the solution.[4]

During the 1980s, game theorists became increasingly disillusioned with this quest, because of two problems. The first is that of *equilibrium selection*. Many of the games that economists want to use as models of the real world have many Nash equilibria, and the assumptions of rationality and common knowledge cannot tell us which of these equilibria will come about. If every player believes that every other player will choose her component of a particular Nash equilibrium, it is rational for each player to choose his component of that equilibrium; but the same applies to each other Nash equilibrium. This problem came to be seen as more daunting as the implications of the Folk theorem sank in: in repeated games, there is typically a vast number of Nash equilibria.

The second problem is the one that was addressed by the *refinement programme*. It is to find a coherent model of what a rational player *would* do, were she to be at a point in a game tree that would not in fact be reached by rational players. It gradually became obvious that, given the presuppositions of classical game theory about what counts as an acceptable solution concept, any adequate solution concept for extensive-form games has to rely on some such model. But it also became obvious that the whole idea of identifying what a rational player would do in a contingency that would occur only if she were not rational was shot through with conceptual problems. Rather than

solving these fundamental problems, the refinement programme seemed to be producing only ad hoc solution concepts supported by casual appeals to intuition.

According to what I take to be the folk history of game theory, the evolutionary turn was welcomed because it offered an escape from these two problems. By introducing a dynamic process to game theory – the dynamic of evolutionary selection – it rationalized the existence of multiple equilibria. (In classical game theory, in which solutions are supposed to be found by the independent deliberations of ideally rational players, it is not clear what it means to speak of a non-unique 'solution'.) Further, by specifying the dynamic process in more detail, game theorists might be able to attack the problem of equilibrium selection in new ways, for example, by developing definitions of stability (thus ruling out unstable equilibria) and measures of the zones of attraction of different equilibria (thus showing which equilibria are most likely to be observed). By legitimizing the relaxation of the assumption of ideal rationality, the evolutionary turn bypassed the roadblock that had defeated the refinement programme.

I think it is significant that the problems that led to the evolutionary turn were theoretical and not empirical. The problem was not that classical game theory generated predictions that were disconfirmed by evidence. The project of classical game theory was not empirical in the first place; the project was to investigate the implications of ideal rationality. Of course, economists were interested in game theory because they hoped to use it to explain the actions of real economic agents. Thus, in the background, there was a presumption that real economic agents are sufficiently like the ideally rational players of the theory for that theory to have some explanatory power. But the development of classical game theory was not significantly constrained by the facts of the world. Disillusionment set in when what was essentially an a priori project ran into internal problems – when it failed to achieve its a priori objectives.

It is also significant that two non-evolutionary ways of tackling these problems were widely known, but not followed up to any significant extent. As long ago as 1960, Schelling (1960) analysed the equilibrium selection problem and showed how, for real human players, equilibrium selection depends on shared conceptions of *prominence* (now usually called *salience*) which allow individuals' expectations to converge on particular equilibria, or *focal points*. Almost all game theorists recognize the force of Schelling's insights, and 'salience' and 'focal point' have become familiar terms in the vocabulary of game theory. However, these concepts have never been integrated into the formal structure of the theory; they are generally used in an ad hoc way, to rationalize intuitively plausible but theoretically unsupported claims about equilibrium selection.[5] Why have these concepts proved so theoretically intractable? A clue to the answer can, I think, be found in Schelling's diagnosis of the limitations of classical game theory. Explaining his analysis of focal points, Schelling insists that the study of games is

'necessarily empirical', because 'the principles relevant to *successful* play, the *strategic* principles, the propositions of a *normative* theory, cannot be derived by purely analytical means from a priori considerations' (1960: 162–3, original emphasis). Twenty years later, in a preface to a reprinting of his seminal book, Schelling expresses his regret that game theory has not developed as he hoped, into an interdisciplinary and applied branch of social science, but has remained 'at the mathematical frontier' (1980: vi). Since game theorists have continued to rely on purely analytical methods, rather than using the empirically-based approaches of sociology and psychology which seem most likely to explain the distinguishing features of focal points, it should be no surprise that they have made little progress in formulating Schelling's insights.

A similar diagnosis of the problems facing the refinement programme was made by Binmore in the mid 1980s. What was needed, according to Binmore (1987), was a theory which could explain the actual reasoning processes of players, which did not assume ideal rationality, and thus which would not be invalidated by players' irrational moves. Thus: 'an attempt must be made to model players' thinking processes explicitly. . . so that deviations from predicted play can be ''explained'' by modifying the model once this has proved inconsistent with observed events' (Binmore 1987: 212). I take it that Binmore means that we need an *empirical* theory of the thinking process actually used by human beings. Clearly, that is what we do need. An a priori theory of ideal rationality cannot tell us what to expect of people who are not ideally rational, while a theory of actual reasoning processes might tell us when to expect lapses from ideal rationality, and what kinds of lapses to expect. The implication, then, is that the refinement programme can be pursued only by using empirical as well as analytical methods – for example, by drawing on the ideas and methods of cognitive psychology. My sense is that most game theorists now acknowledge the validity of Binmore's diagnosis, but – voting with their feet, or with their typing fingers – they have drawn a different implication: that the refinement programme has reached a dead end. Or perhaps, for them, an investigation that can be completed only by empirical work *is* a dead end.

3 THE A PRIORI NATURE OF EVOLUTIONARY GAME THEORY

If the account I have given is correct, the evolutionary turn was a response to *analytical* problems in an essentially a priori research programme, and was taken in preference to other responses which would have required empirical research. But, one might think, the evolutionary approach is empirical too. Is not it one of the most revolutionary features of the evolutionary turn that it substitutes an empirically-based theory of human decision-making for a theory based on ideal rationality?

The answer to this question *ought* to be a clear 'yes': but I want to suggest that the true answer is rather different. As so far practised by economists, evolutionary game theory is predominantly a priori. When game theorists profess to be influenced by biology, they are not thinking of biology as biologists would, as an empirical science struggling to make sense of the facts of the natural world. They are thinking of a small body of mathematical techniques that have proved useful in some of the more theoretical branches of biology.

To support this claim, I shall refer to two recent and major works by evolutionary game theorists: Binmore's two-volume *Game Theory and the Social Contract* (1994, 1998) and Young's *Individual Strategy and Social Structure* (1998). I am conscious of the unfairness of making Binmore and Young the targets of my criticism. These writers stand out among the community of game theorists for their willingness to step outside the mathematical formalisms of game theory, to address foundational questions and to consider some of the big issues of social and political theory. But I have an excuse: it is precisely because Binmore and Young make efforts to explain what their work means that I can use their books as evidence of how evolutionary game theorists interpret their project. The unreflective technicians of the field are more culpable, but by virtue of their lack of reflection, less useful for my present purpose.

Evolutionary game theory, as practised in economics, has certain characteristic features which raise doubts about its empirical seriousness. I shall discuss some of these in the following three sections.

4 ASSUMING THE EXISTENCE OF UTILITY FUNCTIONS

In most evolutionary game theory, the object of analysis is a game. Understood as a formal object, the concept of a game is just as in classical game theory. Thus, players are assumed to have utility functions, which assign cardinal utility indices to outcomes. In addition, it is generally assumed that the mathematical expectation of utility (assessed in terms of objective, frequency-based probabilities) is the relevant measure of a player's success – that is, that at the level of individual decision, the evolutionary process tends to select behaviour which maximizes expected utility. Evolutionary game theory diverges from classical game theory only when it comes to the analysis of beliefs. In the classical theory, players have beliefs about one another which are grounded in, or at least consistent with, ideal rationality and common knowledge; their strategy choices are rational in the sense that they maximize *subjectively* expected utility, when subjective beliefs are themselves rational. Evolutionary game theory does not require the rationality of beliefs.

In some versions of the theory, such as that used by Young, individuals act on beliefs, but those beliefs are formed adaptively from past experience. Thus,

when a game is played repeatedly, each player's beliefs about the behaviour of future opponents tend to track the relative frequencies of similar behaviour on the part of similar opponents in the past. This theoretical approach, sometimes called *fictitious play* or *best-reply dynamics*, does not dispense with rationality altogether, but merely assumes what Young (1998: 144) calls 'low-rationality agents'.

Other versions of evolutionary game theory do not refer to beliefs at all, and consider only the relative frequencies with which different strategies are chosen within the relevant population of players of a game. It is assumed that the rate of change in the frequency with which a particular strategy is played depends on the expected utility it yields (assessed in terms of the actual frequencies of opponents' strategy choices), relative to the expected utility of other strategies. A particular form of this theory, very commonly used in evolutionary game theory, is *replicator dynamics*.[6] The model of replicator dynamics was originally developed for use in biology; its use in economic applications of game theory is often justified by appeal to a biological analogy.

In biology, the *relative fitness* of a phenotype is measured by the expected number of descendants in the next generation of an individual with that phenotype, as a proportion of the expected number of descendants of the phenotype with the highest number of expected descendants. In a sufficiently simplified biological model, in which reproduction is asexual and the properties of each parent's phenotype are perfectly replicated in its offspring, the expected rate of growth of the proportion of the population that has any given phenotype is directly proportional to the relative fitness of that phenotype. If 'expected utility' is substituted for 'expected number of descendants', and if it is assumed that higher values of expected utility lead to increases in the frequency of the associated behaviour in something like the same way that higher expected numbers of descendants lead to increases in the frequency of the parental genotype, we arrive at the economic version of replicator dynamics.

I suggest that these strategies for retaining expected utility theory within evolutionary game theory are opportunistic. In classical game theory, the assumption that utility functions exist is grounded in a theory of rational choice. (Recall that von Neumann and Morgenstern produced their axiomatic formulation of expected utility theory to persuade a sceptical economics profession of the meaningfulness of the cardinal utility indices that game theory needs.) Since classical game theory is a theory of ideal rationality, it is wholly appropriate that it should depend on axioms of rational preference. (Whether the axioms required by classical game theory really are rationally compelling is another matter.) But *evolutionary* game theory cannot make the same appeal.

When I say that evolutionary game theorists are being opportunistic, what I mean is this. In mathematical terms, the project of classical game theory was

to formulate solution concepts for games. This project ran into analytical problems; but those problems did not call into question the theoretical coherence of the concept of a game. In particular, they did not call into question the theoretical coherence of utility indices. Seen as a purely mathematical enterprise, the project could be continued by introducing dynamic processes, isomorphic with those used by some theoretical biologists, while retaining the original concept of utility. The tension between appealing to the axioms of expected utility theory to legitimate the definition of a game and rejecting rationality at the level of beliefs is relevant only to the *interpretation* of the project. My suspicion is that, for many game theorists, interpretation is a secondary matter: internal contradictions in their interpretations of different elements of their theories do not particularly bother them.

From this final charge, I exempt both Young and Binmore. Young is quite clear about what he is doing in dropping some of the rationality assumptions of game theory while keeping others. He obviously has in mind some model of 'low-rationality agents' whose preferences can be represented as expected utility maximizing, but whose beliefs are not necessarily rational in the Bayesian sense. Nevertheless, Young does not try to develop any such model. Nor does he do much to persuade his readers that such a model would be psychologically credible or predictively successful. Of course, no individual theorist can be expected to do everything that is needed to complete the project on which he is working. But the *general* lack of interest among evolutionary game theorists in developing and defending such a model does, I think, suggest that something is awry.

Binmore is much more ambitious. He does not appeal to rationality at all, but only to evolutionary selection (interpreted broadly, so as to include learning by experience and imitation as well as genetic evolution and biological natural selection). Thus, he says, if it is valid to model people as maximizers, this can only be because 'evolutionary forces, biological, social, and economic, [are] responsible for getting things maximized' (1994: 20). And he claims that it *is* valid to model people as maximizers of expected utility: that people act as if maximizing expected utility is itself a product of evolutionary selection. My impression, again gathered from informal discussions, is that evolutionary game theorists often use this argument as a first line of defence when challenged to justify their use of expected utility theory. Most of them, I suspect, have not thought seriously about whether the argument really works; they simply have a hunch that it could be made to work, and are satisfied with that. Again, it is worrying that so little work has been done to develop and test a claim that is so important for the validity of evolutionary game theory. Binmore, however, has tried to present the argument explicitly. I shall discuss Binmore's analysis in the next section.

5 NATURAL SELECTION AS A TAUTOLOGY

Interpreted at a sufficiently general level, Darwin's theory of natural selection can be read as a tautology. Consider any population made up of different types of objects. Suppose that each object is capable of replicating itself in some way. (Think of the survival of an object from one period to the next as a special case of replication.) If different types of object replicate at different rates, the relative frequencies of those types of object in the population will change over time; the types with the higher rates of replication will become relatively more frequent. This is not an empirical hypothesis, it is an analytical truth. So if 'fitness' means no more than 'rate of replication', the principle of 'the survival of the fittest' can be presented as a tautology. I do not mean to imply that the principle, understood in this way, is trivial. It is an important truth, and understanding it is an essential first step in understanding evolutionary processes. But, I suggest, evolutionary game theorists are inclined to over-estimate the usefulness of the principle, when it is not supplemented by empirical assumptions.

Binmore's defence of the assumption of expected utility maximization provides a good illustration. He begins by appealing to Dawkins's (1976) concept of a *meme*, which Binmore defines as 'a norm, an idea, a rule of thumb, a code of conduct – something that can be replicated from one head to another by imitation or education, and that determines some aspects of the behavior of the person in whose head it is lodged'. A process analogous with biological natural selection selects those memes that are most successful at replicating themselves. In this model, people are not assumed to be rational in any conscious sense. They feature in the model merely as hosts for memes; it is as if they are *infected* by the memes they carry. But in equilibrium, and to an outside observer, 'it will seem as though the infected agent is acting in his own self-interest, provided that the notion of self-interest is interpreted as being *whatever makes the bearer of a meme a locus for replication of the meme to other heads*' (1994: 20, original emphasis). From this conclusion, Binmore goes on to argue that, at the level of individuals' preferences, the process of meme selection will induce preferences that satisfy the kinds of consistency conditions that are conventionally assumed in economic theory:

> . . . the practical reasons for thinking consistency an important characteristic of a decision-maker cannot be lightly rejected. People who are inconsistent will necessarily sometimes be wrong and hence will be at a disadvantage compared to those who are always right. And evolution will not be kind to memes that inhibit their own replication. (1994: 27)

I have reproduced all the essential steps in Binmore's argument. Notice that it does not appeal to any empirical facts, or make any empirical assumptions. A meme, in this argument, is an entirely abstract entity. All that we asked to assume about memes is that they exist in some unspecified

form in the human brain, that they govern individual behaviour in some unspecified way, and that they can replicate from brain to brain by some unspecified mechanism. Using the principle of natural selection in its purely tautological form, Binmore claims to derive a substantive empirical conclusion, that the behaviour of human individuals will tend towards a pattern that can be rationalized by consistent preferences.

But how can such an argument possibly work? Empirical conclusions cannot be derived from non-empirical premises. The error occurs when Binmore moves from propositions that are true for memes to propositions that are true for people. He is entitled to conclude that, in equilibrium, each *meme* that exists is 'behaving' as if maximizing *its* expected rate of replication. Each person's behaviour is then whatever is induced by the interaction of the collection of memes that she carries. But unless we assume that each person carries only one meme – which is hardly credible, given Binmore's interpretation of a meme – we cannot infer that the person's behaviour maximizes anything. For the same reason, Binmore's concept of 'right' and 'wrong' behaviour is ill-defined. In the context, 'right' seems to mean 'maximizing the replication of memes'; but which memes?

In relying so heavily on a priori reasoning about natural selection, evolutionary game theorists are not following the model of biology. Natural selection is only one element of the biological theory of evolution. It is a commonplace to say that modern evolutionary biology is a synthesis of Darwin's theory of natural selection and Mendel's theory of genetics. The theory of genetics explains the process of biological replication: it explains *what* gets replicated (genes, DNA sequences) and *how* replication takes place. Natural selection does not directly select phenotypes, still less patterns of behaviour, that tend to favour the survival of the relevant individual or to enhance its prospects for engaging in reproduction. It selects *genes* that are successful *in replicating themselves*. In order to understand the implications of natural selection at the level of phenotypes, we have to understand how phenotypes are related to genes, and how genes replicate.

For example, natural selection working at the level of genes is quite compatible with there being a stable proportion of unfit phenotypes in the population. This is because the reproductive success of a phenotype depends on the *combination* of genes that it carries. A gene that is beneficial in some combinations may be harmful in others. Thus, a gene pool that is in equilibrium may contain genes which, when brought together in the same individual by the random processes of sexual reproduction, have disastrous consequences for that individual's prospects of survival and reproduction. In consequence, it is not universally true that natural selection favours plants and animals that are successful at survival and reproduction. Perhaps this is approximately true, but if it is, it is so by virtue of particular facts about the biological world, which have been discovered only by empirical investigation.

One highly relevant and very special feature of genes is that the genes carried by an individual organism normally replicate *together*. In a sexually-reproducing species, each individual passes on a random selection of his or her genes in each episode of reproduction. In consequence, we might say, the genes carried by an individual have many interests in common: in order for any gene to replicate, the individual who carries it has to be involved in an episode which gives every one of his or her genes a chance to replicate. If genes were capable of reproducing independently of one another, as memes may very well be able to do, the behaviour of biological organisms would surely show less apparent coherence and purposefulness than they in fact do. (Binmore's analogy of infectious diseases illustrates this point. Many agents of infection are able to replicate independently of the replication of their hosts' genetic inheritance; they can induce states in their hosts – terminal illness, for example – which do not look at all like purposefulness at the level of the host.)

The evolutionary processes represented in economic applications of evolutionary game theory are not (or are not primarily) biological. They are processes of human learning and imitation. Thus, economists cannot piggyback on biologists' understanding of genetics. We need theories which tell us, for the relevant processes of learning and imitation, what gets replicated and how it gets replicated. It is only when we know this that we can legitimately make use of the tautologies of natural selection. Merely to *assume* the existence of utility functions, and to *assume* replicator dynamics (or some similar dynamic process), is to try to create an evolutionary theory that is analogous with a biology that understands natural selection but knows nothing about genetics.

Within evolutionary game theory, surprisingly little work has been done to investigate how imitation and learning actually work, or even (a project which might be more congenial to game theorists) how these processes *might* work. But such work as has been done suggests that, starting from a psychologically credible model of imitation or learning, expected utility theory does *not* emerge as the natural outcome of selection. By assuming very special values of the parameters of the process of imitation or learning, it is *possible* to derive expected utility theory, but there seems to be no good reason to suppose, as a matter of empirical fact, that those special values occur.[7]

It might seem that what is needed is a further tier of evolutionary argument, to show that the heuristics that human beings use in imitation and learning are fine-tuned by some selection process so that the parameters of these processes come to have the 'right' values. Can not we use the analogy of animal learning? There are good evolutionary reasons to expect that the heuristics that animals use in learning are fine-tuned to maximize expected reproductive success in natural environments. But the analogy is flawed. In the case of animal behaviour, we *already know* what is being selected for, namely reproductive success, and so we are entitled to infer that selection will

favour learning heuristics that serve reproductive success. But whether the economic environment contains forces that select expected utility maximizing behaviour is an open question, which we are trying to resolve by theorizing about the processes of imitation and learning. We are entitled to infer that selection will favour learning heuristics which induce whatever kinds of behaviour tend to replicate those heuristics; but we are not entitled to assume that such behaviour is expected utility maximizing.

6 EXPUNGING HISTORY

I have said that it is a commonplace that evolutionary biology is a synthesis of Darwinian and Mendelian theories. Perhaps less often said, but still true, is that there is a third strand to evolutionary biology: the study of the actual contingencies of evolutionary history. I take it that the purpose of a scientific theory is to explain regularities in our observations of the world. The biological world contains many regularities that can be explained only by referring to historical contingencies.

For example, there are very great similarities in anatomy and behaviour between human beings and chimpanzees. This is an empirical regularity that can be used (and indeed is used, in psychological and medical research) to make reliable predictions about one species from observations of the other. If we looked only at the environments in which the two species now live, we would find little reason to expect such similarities to have been the result of natural selection; we might expect there to be much closer resemblances between chimpanzees and (say) capuchin monkeys than between chimpanzees and humans. To explain the similarity between chimpanzees and humans, we have to reconstruct evolutionary history, and discover that the two species have a relatively recent common ancestor. Since the species diverged, there has not been enough time for large differences between them to accumulate, despite the enormous differences between the environments in which they now live. The case of marine mammals provides another example: the fact that in many respects dolphins are more like land-based mammals than they are like fish can be explained only by reconstructing the common ancestry of all mammals. Or again: the remarkable regularities in the geographical distribution of placental and marsupial mammals were explained only when continental drift was understood. That marsupials are found in Australia and South America but not Europe is a consequence of the historical fact that, relatively recently, Australia and South America were parts of a single continent. Notice that in all these cases, historical contingencies are not being used to explain one-off events, in the way that historians explain one event in terms of another; they are essential components of explanations of very general empirical regularities.

It is, I suggest, a crucial feature of evolutionary explanation that historical contingencies are important. If one thing grows out of another, then to

explain the properties of the descendant we need to consider the properties of the ancestor. Thus, an evolutionary approach to social science must take account of the actual facts of history. As an example of what such a social science might look like, consider linguistics. Linguists study the evolution of languages. They explain similarities between languages (for example, that Finnish is much more similar to Hungarian than it is to Swedish) by reconstructing the history of language evolution. Some major regularities in language use turn out to be the product of one-off historical events (for example, the fact that English has many more French-derived words than Dutch does is largely the result of the success of William the Conqueror at the Battle of Hastings in 1066). Again, it is important to recognize that linguistics is at least as much a science as economics is: it is engaged in explaining linguistic regularities, not just recording linguistic facts.

If I am right, an evolutionary approach to economics must be historical. But where is the history in evolutionary game theory, as this is currently practised? Where, even, is the *space* in the theory into which the influences of historical contingency could be slotted? As an example of what I have in mind in posing the latter question, think of Schelling's concept of salience. Everyone recognizes that what is salient to particular human beings is in large part a product of their particular social experiences; thus, criteria of salience can be expected to have evolutionary histories, in much the same way that languages do. (Indeed, if languages are understood as solutions to coordination problems, languages *are* focal points in the Schelling sense.) Thus, we might expect that an adequate evolutionary game theory would include a theory of how ideas of salience evolve. But for that to be possible, the framework of evolutionary game theory must at least allow salience to be represented. By taking as its subject matter games, as conventionally defined (and thus without taking any account of how different strategies are 'labelled' – that is, of how they are conceptualized by the players themselves), evolutionary game theory allows no space for salience in its explanations.[8]

Evolutionary game theorists avoid having to confront the facts of history by confining themselves to the a priori analysis of natural selection within the artificial domain of formal games. They look for 'results' which can be shown to hold, irrespective of historical contingencies. Since in the short run, the state of an evolutionary system at any given time depends on where it was previously, the search for such results tends to lead to a focus on very long-run equilibrium properties. These, I take it, have to be understood as a way of describing general tendencies that are constantly operating in the real world, even though (in the long-run sense) that world is in perpetual disequilibrium. This approach can lead evolutionary game theorists to investigate tendencies which are so weak, or work only over such very long expanses of time, as to have little useful explanatory power.

Young's (1993, 1998) work on 'stochastic stability' illustrates this assertion. Explaining his approach, Young notes that most evolutionary game

theorists have omitted stochastic disturbances from their models and examined the *expected* motion of evolutionary processes as if those processes were deterministic. Young's strategy is to build a 'noise term' into his models of dynamic processes. What are the advantages of this strategy? Young:

> While [the expected motion of a stochastic process] may be a reasonable approximation of the short-run (or even medium-run) behavior of the process, however, it may be a very poor indicator of the long-run behavior of the process. A remarkable feature of stochastic dynamical systems is that their long-run (asymptotic) behavior can differ radically from the corresponding deterministic process *no matter how small the noise term is* . . . But there is also an unexpected dividend: [the] *long-run average* behaviour [of such processes] can be predicted much more sharply than that of the corresponding deterministic dynamics, whose motion usually depends on the initial state. (1998: 47, original emphasis)

Roughly, what Young means is this. In a deterministic dynamic model, there may be more than one stable equilibrium. Thus, we cannot predict which equilibrium will come about unless we know the initial state of the system. But if we add a small disturbance term to the model, no equilibrium can persist indefinitely: sooner or later, a coincidence of random disturbances will shift the system out of the basin of attraction of one equilibrium and into that of another. Thus, in a typical model, there are long periods of apparent stability, separated by brief periods of rapid transition from one almost stable state to another. Young focuses on the very long-run properties of a system: if the system is observed for an indefinite length of time, for what proportion of that time can it be expected to be in each state? If the extent of stochastic variation is sufficiently small, then many systems have the property that they can be expected to spend almost all of that indefinite expanse of time in a particular state. Such a state is *stochastically stable*. Young investigates whether, for various classes of games, there are stochastically stable states. He uses this method to 'recover' various solution concepts which have been proposed in classical game theory. That is, he shows that the strategy combinations picked out by those solution concepts correspond with stochastically stable states in evolutionary models: 'Interpreted in this way, the evolutionary approach is a means of reconstructing game theory with minimal requirements about knowledge and rationality' (1998: 144).

Notice that what Young is doing is to expunge historical contingency from his models by considering a 'long run' in which the effects of such contingencies cancel out. His account of what he is doing conveys the strong hint that the *point* of studying this long run is to arrive at 'sharp' predictions, and that there is something *unsatisfactory* about models in which the motion of the system depends on the initial state. This way of thinking maintains, in the supposedly empirical realm of evolution, the classical game theorists' search for the Holy Grail of a unique solution for every game.

But if our object is to explain the world as it is, the only justification for excising history-dependence is that the regularities that we are trying to understand *really are* history-independent. I recognize that Young's approach does throw useful light on some of the dynamic processes of the real social world. (For example, it fills in some of the gaps in Schelling's famously insightful model of the evolution of patterns of racial segregation.)[9] But in many of Young's models, the long run seems to be extraordinarily long, perhaps even a matter of billions of years. (How long would it take Britain to switch to driving on the right if we waited for a coincidence of random mistakes by individual drivers? And that would be just *one* transition between equilibria: Young's long run is a period which contains a very large number of transitions.[10]) Young's analysis, I suggest, illustrates a persistent tendency of evolutionary game theorists to look for conclusions that can be established by a priori methods, rather than to accept the challenge of explaining the regularities in the world about us.

7 SLASH AND BURN

I have suggested that the evolutionary turn in game theory may not be so revolutionary after all. If game theorists' claims are taken at face value, the theory of human behaviour that underlies the evolutionary approach is fundamentally different from that which is used in classical game theory. But in fact, large parts of the old theory have been imported into the new framework, with little attempt to check that their validity is preserved when the interpretations given to their concepts are changed. What we see is not so much a challenge to the theory of rational choice as a superficial restyling of it.

It is worth reflecting on the fact that the theory of rational choice went through a previous restyling operation, just about a century ago. Recall that the first neoclassical economists, working in the late nineteenth century, conceived of rational choice in instrumental terms: they postulated that, for each individual, there is a one-dimensional measure of psychological satisfaction, 'utility', and they defined rationality as the maximization of utility. These assumptions allowed economists to make use of the powerful mathematical tools of constrained maximization, and to construct elegant general theories from simple premises. This mode of theorising was essential a priori, but was based on empirical assumptions about human psychology. Unfortunately, however, the postulated one-dimensional measure of satisfaction had not been found (and, a century later, it remains unfound). This left an embarrassing gap in economists' explanations of the world.

One might have thought that economics, as a scientific discipline, would have tried to fill this gap by empirical research, by investigating the facts of human psychology and then by revising the theory as necessary to fit those facts as they were discovered. Instead, in a move that was initiated by Vilfredo Pareto and that culminated in Paul Samuelson's revealed preference theory

and in Leonard Savage's formulation of expected utility theory, economics retained the mathematical structure of utility maximization while throwing away the psychological assumptions that had originally been thought to justify it. In place of instrumental rationality, economics substituted a wholly a priori theory of the internal consistency of preferences. But this radical change at the level of foundations was carried out without major changes to the superstructure of the theory. One is entitled to be sceptical about such a manoeuvre: it prompts the suspicion that what are supposed to be foundations are really just cladding.[11] And now, we are seeing *another* supposedly fundamental change in the foundations of economic theory; and again, the superstructure is being preserved.

For me, reflecting on this history induces a sense of unease about the seriousness of the whole enterprise. What seems to be revealed is an endemic unwillingness on the part of economic theorists of decision-making to face up to empirical questions. It seems that the most persistent feature of the theory is not any unifying explanatory principle, but a commitment to an a priori mode of enquiry. It is as if, when a line of research runs into a fundamental problem that can be solved only by empirical research, that line has to be closed down – as if it is better to conserve the formal structure of a theory and to give it a new interpretation than to conserve the questions that are being asked and to look for new ways of answering them. The result is a slash and burn approach to social science – an approach which uses a priori methods to derive quick results in whatever is the current field of investigation, then moves on as soon as real empirical work become necessary. My fear is that the evolutionary turn in game theory will turn out to be just another episode of slash and burn.

But I do not want to be too pessimistic. Evolutionary game theory is still in its infancy. A *genuinely* evolutionary approach to economic explanation has an enormous amount to offer; biology really is a much better role model for economics than is physics. I just hope that economists will come to see the need to emulate the empirical research methods of biology, and not just its mathematical techniques.

Robert Sugden
University of East Anglia

NOTES

1 From now on, I shall use the term 'game theory' as a shorthand for 'game theory as used in economics', and 'game theorist' as a shorthand for 'economic theorist who uses game theory as a principal tool'. These shorthands should not be read as parochialisms: this essay is concerned only with the use of evolutionary ideas *in economics*.

2 I shall not divulge the game theorists's name; revealing his sex will do little to narrow down the field of suspects.

3 For methodological perspectives on this debate, see Hausmann (1992, Chapter 13), Tammi (1999), and Guala (2000).

4 The theorists who have come closest to finding the Holy Grail are Harsanyi and Selten (1988).
5 There have been a few attempts to accommodate salience within the formal structure of a theory of rational choice; see, e.g. Bacharach (1993) and Sugden (1995).
6 The model of replicator dynamics was introduced by Taylor and Jonker (1978).
7 These remarks are informed by the theoretical work reported by Börgers and Sarin (1996, 1997) and Cubitt and Sugden (1998). Interested readers should consult both of the versions of Börgers and Sarin's paper; some of the more discouraging conclusions reached in the earlier version are downplayed in the later one.
8 I discuss this issue more fully in Sugden (1998).
9 See Young (1998: 6–10, 62–5); the original model is in Schelling (1978: 137–66).
10 In a rare historical aside, Young (1998: 16–17) summarizes the actual history of 'keep left' and 'keep right' conventions in Europe, and claims that this exhibits the patterns predicted by his theory. But he has to use a model in which nations, not individuals, are players; the French Revolution counts as a single exogenous shock.
11 I explain the reasons for my suspicion in Sugden (1991).

REFERENCES

Bacharach, Michael (1993) 'Variable universe games', in Ken Binmore, Alan Kirman and P. Tani (eds) *Frontiers of Game Theory*, Cambridge, MA: MIT Press, pp. 255–75.
Binmore, Ken (1987) 'Modeling rational players: Part I'. *Economics and Philosophy* 3: 179–214.
Binmore, Ken (1994) *Game Theory and the Social Contract*, Volume I: *Playing Fair*, Cambridge, MA: MIT Press.
Binmore, Ken (1998) *Game Theory and the Social Contact*, Volume II: *Just Playing*, Cambridge, MA: MIT Press.
Börgers, Tilman and Rajiv Sarin (1996) Naive reinforcement learning with endogenous aspirations. Mimeo, University College London.
Börgers, Tilman and Rajiv Sarin (1997) 'Learning through reinforcement and replicator dynamics', *Journal of Economic Theory* 77: 1–14.
Cubitt, Robin P. and Robert Sugden (1998) 'The selection of preferences through imitation', *Review of Economic Studies* 65: 761–71.
Dawkins, Richard (1976) *The Selfish Gene*, Oxford: Oxford University Press.
Guala, Francesco (2000) 'The logic of normative falsification: rationality and experiments in decision theory', *Journal of Economic Methodology* 7: 59–93.
Harsanyi, John C. and Reinhard Selten (1988) *A General Theory of Equilibrium Selection in Games*, Cambridge, MA: MIT Press.
Hausman, Daniel M. (1992) *The Inexact and Separate Science of Economics*, Cambridge: Cambridge University Press.
Schelling, Thomas (1960) *The Strategy of Conflict*, Cambridge, MA: Harvard University Press, Second edition 1980.
Schelling, Thomas (1978) *Micromotives and Macrobehavior*, New York: Norton.
Sugden, Robert (1991) 'Rational choice: a survey of contributions from economics and philosophy', *Economic Journal* 101: 751–85.
Sugden, Robert (1995) 'A theory of focal points', *Economic Journal* 105: 1269–302.
Sugden, Robert (1998) 'The role of inductive reasoning in the evolution of conventions', *Law and Philosophy* 17: 377–410.

Tammi, Timo (1999) 'Incentives and preference reversals: escape moves and community decisions in experimental economics', *Journal of Economic Methodology* 6: 351–80.
Taylor, P. and L. Jonker (1978) 'Evolutionary stable strategies and game dynamics', *Mathematical Biosciences* 40: 145–56.
Young, H. Peyton (1993) 'The evolution of conventions', *Econometrica* 61: 57–84.
Young, H. Peyton (1998) *Individual Strategy and Social Structure*, Princeton, NJ: Princeton University Press.