Journal of Consciousness Studies
www.imprint-academic.com/jcs

Brian Skyrms

Game Theory, Rationality and Evolution of the Social Contract

Game theory based on rational choice is compared with game theory based on evolutionary, or other adaptive, dynamics. The Nash equilibrium concept has a central role to play in both theories, even though one makes extremely strong assumptions about cognitive capacities and common knowledge of the players, and the other does not. Nevertheless, there are also important differences between the two theories. These differences are illustrated in a number of games that model types of interaction that are key ingredients for any theory of the social contract.

Introduction

The Theory of Games was conceived by von Neumann and Morgenstern as a theory of interactive decisions for ideally rational agents. So was the theory of the social contract, from the Sophists to Thomas Hobbes. Hobbes wanted to bring the rigor and certainty of Euclidean geometry to social philosophy. If he fell somewhat short of this goal, even by the standards of his own time, perhaps the theory of games could be utilized to complete the project. This idea has been pursued in different ways by John Harsanyi, John Rawls, David Gauthier and Ken Binmore.

Rational Choice

But the foundation of the classical theory of games on the theory of rational decision has itself proved more complicated than it seemed at the time. Von Neumann and Morgenstern thought of the theory of rationality as work in progress which, when complete, would specify a unique correct rational act for any decision maker in any decision situation. If such a theory of rationality were in hand, then it appears that Nash equilibrium would be the unique outcome of the decisions of rational agents in an interactive decision situation. A *Nash equilibrium* is defined as a specification of an act for each agent, such that no agent can gain by unilateral change of her act. Now, the argument goes, each decision maker can figure out what the other decision makers will do from the theory of rationality. Thus, the only state consistent with rationality is that each player maximizes payoff given knowledge of what the other players do —

a Nash equilibrium. You can find this argument applied within the theory of two-person zero-sum games — within which it makes a good deal of sense — in *The Theory of Games and Economic Behavior* (von Neumann & Morgenstern, 1947), p. 148.

There are some tacit assumptions to the argument. In the first place, it is not only assumed that each agent is rational, but also that every agent knows so — in order to deduce the actions of the others from the theory. But, if the theory is predicated on all the agents being able to deduce the actions of the others, then agents must not only know that others are rational, but also know that others know that they are — otherwise others might not use the theory, and so forth. That is, we assume not only that all the agents are rational, but also that rationality is *common knowledge*. Likewise we must assume that the agents correctly identify the decision situation to which they are applying the theory, and that they know that all others do so as well, and that they know that all others know this, and so forth.

In the second place, it is assumed that the theory of rationality singles out a unique correct act. This is 'almost' true in the theory of zero-sum games. Where multiple alternatives are allowed, they are interchangeable in an appropriate sense. But it is wildly false when we move into the territory of non-zero sum games. For instance, consider the following co-ordination game. Two players each independently pick from a list of three colours. If they pick the same, they win a prize; otherwise they lose. The *symmetry* of the situation precludes any reasonable unique prescription by any theory of rational choice.

Since the payoffs are invariant under permutations of colours and the colours themselves are assumed to have no significance other than as labels, the strategy that is uniquely recommended must also be invariant under permutations of colours. The only Nash equilibrium invariant in this way consists in each player independently randomizing between the three colours with equal probabilities. This is a unique prescription, but hardly a reasonable one since it leads the players to mis-coordinate two-thids of the time. This is the Curse of Symmetry that plagues theories that, like Harsanyi and Selten (1988), pursue the goal of a uniquely selected rational act for each player.

Once the hope of a theory of rationality that always singles out a unique rational act is given up, the von Neumann-Morgenstern justification of Nash equilibrium unravels. Perhaps I choose Red, thinking that you will choose Red and you choose Green, thinking that I will choose Green. We are acting rationally, given our beliefs, but we mis-coordinated and are not at a Nash equilibrium.

It is time to back up and see what sort of results we can get from assumptions of rational choice. To do so, we cannot continue to be vague about the term 'rationality'. We say that, for a given player, act B is *weakly dominated* by act A, if, no matter what the other players do, act A leads a payoff at least as great as act B does, and for *some* combination of acts by other players, act A leads to a greater payoff than act B. We say that act A *strongly* dominates act B if, for *all* combinations of acts by other players, act A leads to a greater payoff than act B. We say that a player is *Bayes-rational* if she acts so as to maximize her expected payoff, given her own degrees of belief and her own evaluation of consequences. If a player is *Bayes-rational*, she will not choose a *strongly dominated* act because the act that strongly dominates it must have greater expected payoff, no matter what her degrees-of-belief about what other players will do. If a player is *Bayes rational* and, in addition, gives every combination of other players' actions non-zero

probability, then she will not choose a *weakly dominated* act, because in this case the act that weakly dominates will have greater expected payoff.

Suppose that the setting for rational choice based game theory is one where not only are all the players Bayes-rational, but furthermore it is common knowledge among the players that they are all Bayes-rational. Assume also that the game being played is also common knowledge. What does this tell us about the outcomes? Each player's play must maximize expected payoff given her beliefs about others players' play, and each player's beliefs about other players' beliefs and play must be consistent with them maximizing expected payoff, and each player's beliefs about other players' beliefs about her beliefs and her play must be consistent with her maximizing expected payoff, and so forth to arbitrary high levels. Bernheim (1984) and Pearce (1984) investigate strategies that are consistent with common knowledge of rationality, which they call *rationalizable* (or 'correlated rationalizable' in the general case). They show that such strategies are those that remain after *iterated deletion of strongly dominated strategies*. (You identify all the dominated strategies for all the players and delete them. After they are deleted, other strategies may become strongly dominated. Then you delete those, and so forth until you come to a stage where there are no strongly dominated strategies to delete.)

Common knowledge of Bayesian rationality imposes much weaker constraints on play than Nash equilibrium. In our co-ordination game, *any* profile of play is consistent with common knowledge of Bayesian rationality. But in some situations, it leads to a unique outcome. Consider the following 'Beauty Contest' game (Moulin, 1986; Nagel, 1995). A finite number of people play. Each names an integer from 0 to 100. The person who is closest to half the mean wins a prize, others get nothing. (In case of a tie, the prize is shared.) In the beauty contest game, the prize is independent of the actions of the players.[1] The unique outcome consistent with common knowledge of Bayesian rationality is that each person chooses zero.

First, notice that the highest that half of the mean could be (if everyone chose 100) is fifty, so it makes no sense choosing a number greater than fifty. But everyone knows this, so no one will choose more than fifty, in which case it makes no sense choosing a number greater than twenty-five. Iterate this argument, and any number greater than zero gets eliminated as a rational choice. This shows that iterated deletion of weakly dominated strategies eliminates everything but zero. A slightly more complex argument using mixed strategies can be given to show that iterated elimination of strongly dominated strategies eliminates everything but zero. Everyone choosing zero is the unique outcome of this game consistent with common knowledge of Bayesian rationality. If you try this experiment in a room full of people, no matter how sophisticated, you will not get the result predicted by common knowledge of rationality. This is well-documented in the experimental literature (Nagel, 1995; Ho,Weigelt and Camerer, 1996; Stahl, 1996; Camerer, 1997; Duffy and Nagel, 1997).

These are called 'Beauty Contest' games because they use the kind of iterated reasoning that Keynes described in a famous analogy to the stock market:

> . . . professional investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the

[1] This sentence was added during peer commentary in response to a query from Herbert Gintis.

> average preferences of the competitors as a whole . . . . It is not a case of choosing which, to the best of one's judgement, are really the prettiest, nor even those which average opinion thinks are the prettiest. We have reached the third degree where we devote our intelligences to anticipating what the average opinion expects the average opinion to be. There are some, I believe, who practice the fourth, fifth and higher degrees (pp. 155–6).

Common knowledge of rationality requires arbitrarily high degrees of this reasoning.

Is it so surprising that people typically fail to satisfy the assumption of common knowledge of Bayesian rationality? Even to state the assumption requires an infinite hierarchy of degrees of belief, degrees of belief over degrees of belief, and so on. The model can be shown to be mathematically consistent, but it is an abstraction far removed from reality. One might retreat from the assumption of common knowledge of Bayesian rationality and try to base game theory on the simple assumption that the players *are* Bayesian rational. The resulting theory is so weak as to be hardly worth pursuing. It says that players will not choose strongly dominated acts — acts with an alternative that carries a better payoff better no matter what the other players do. (This assumption itself is difficult to square with some experimentally observed behaviour.) One is reminded of Spinoza's characterization of his predecessors: ' . . . they conceive of men, not as they are, but as they themselves would like them to be . . . '.

Adaptive Dynamics

Hobbes wanted a theory of the social contract based in self-interested rational choice. David Hume represents a different tradition. For Hume, the social contract is a tissue of conventions which have grown up over time. I cannot resist reproducing in full this marvellously insightful passage from his *Treatise:*

> Two men who pull on the oars of a boat do it by an agreement or convention, tho' they have never given promises to each other. Nor is the rule concerning the stability of possession the less deriv'd from human conventions, that it arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it. On the contrary, this experience assures us still more, that the sense of interest has become common to all our fellows, and gives us a confidence of the future regularity of their conduct: And 'tis only on the expectation of this, that our moderation and abstinence are founded. In like manner are languages gradually establish'd by human conventions without any promise. In like manner do gold and silver become the common measures of exchange, and are esteem'd sufficient payment for what is of a hundred times their value (p. 490).

Hume is interested in how we actually got the contract we now have. He believes that we should study the processes that lead to a gradual establishment of social norms and conventions. Modern Humeans, such as Sugden, Binmore, Gibbard, take inspiration as well from Darwinian dynamics. The social contract has evolved, and will continue to evolve. Different cultures, with their alternative social conventions, may be instances of different equilibria, each with its own basin of attraction. The proper way to pursue modern Humean social philosophy is via dynamic modelling of cultural evolution and social learning.

In Evolutionary theory, as in classical game theory, there is strategic interaction in the form of frequency-dependent selection, but there is no presumption of rationality — let alone common knowledge of rationality — operative. Indeed the organisms that are evolving may not even be making decisions at all. Nevertheless, game

theoretic ideas have been fruitfully applied to strategic interaction in evolution. And the key equilibrium concept is almost the same.

The most striking fact about the relationship between evolutionary game theory and economic game theory is that, at the most basic level, a theory built of hyper-rational actors and a theory built of possibly non-rational actors are in fundamental agreement. This fact has been widely noticed, and its importance can hardly be over-estimated. Criticism of game theory based on the failure of rationality assumptions must be reconsidered from the viewpoint of adaptive processes. There are many roads to the Nash equilibrium concept, only one of which is based on highly idealized rationality assumptions.

However, as we look more closely at the theory in the evolutionary and rational choice settings, differences begin to emerge. At the onset, a single population evolutionary setting imposes a symmetry requirement which selects Nash equilibria which might appear implausible in other settings. Furthermore, refinements of the Nash equilibrium are handled differently. Standard evolutionary dynamics, the replicator dynamics, does not guarantee elimination of weakly dominated strategies. In a closely related phenomenon, when we consider extensive form games, evolutionary dynamics need not eliminate strategies which fail the test of sequential rationality. Going further, we shall see that if we generalize the theories to allow for correlation, we find that the two theories can diverge dramatically. Correlated evolutionary game theory can even allow for the fixation of strongly dominated strategies. These are strategies which fail under even the weakest theory of rational choice — the theory that players are in fact Bayes rational.

The situation is therefore more complicated than it might at first appear. There are aspects of accord between evolutionary game theory and rational game theory as well as areas of difference. This is as true for cultural evolution as for biological evolution. The phenomena in question thus have considerable interest for social and political philosophy, and touch some recurrent themes in Hobbes and Hume.

Evolutionary Game Theory

Let us consider the case of individuals who are paired at random from a large population to engage in a strategic interaction. Reproduction is asexual and individuals breed true. We measure the payoffs in terms of evolutionary fitness — expected number of offspring. Payoff to a strategy depends on what strategy it is paired against, so we have frequency dependent selection in the population. We write the payoff of strategy A when played against strategy B as U(A|B). Under these assumptions, the expected Fitness for a strategy is an average of its payoffs against alternative strategies weighted by the population proportions of the other strategies:

$$U(A) = SUM_i\, U(A|B_i)\, P(B_i)$$

The expected fitness of the population is the average of the fitnesses of the strategies with the weights of the average being the population proportions:

$$UBAR = SUM_j\, U(A_j)\, P(A_j)$$

We postulate that the population is large enough that we may safely assume that you get what you expect.

The population dynamics is then deterministic. Assuming discrete generations with one contest per individual per generation, we get:

DARWIN MAP: P′(A) = P(A) U(A)/UBAR

where P(A) is the proportion of the population today using strategy A and P′ is the proportion tomorrow. Letting the time between generations become small, we find the corresponding continuous dynamics:

DARWIN FLOW: dP(A)/dt = P(A) [U(A)-UBAR]/UBAR

The Darwin flow has the same orbits as the simpler dynamics obtained by discarding the denominator, although the velocity along the orbits may differ:

REPLICATOR FLOW: dP(A)/dt = P(A) [U(A)-UBAR]

In the following discussion, we will concentrate on this replicator dynamics. It is of some interest that the replicator dynamics emerges naturally from a number of different models of cultural evolution based on imitation (Binmore, Gale and Samuelson, 1995; Bjornerstedt and Weibull, 1995; Sacco, 1995; Samuelson, 1997; Schlag, 1994; 1996).

The replicator flow was introduced by Taylor and Jonker (1978) to build a foundation for the concept of evolutionarily stable strategy introduced by Maynard-Smith and Price (1973). The leading idea of Maynard-Smith and Price was that of a strategy such that if the whole population uses that strategy, it cannot be successfully invaded. That is to say, that if the population were invaded by a very small proportion of individuals playing a different strategy, the invaders would have a smaller average payoff than that of the natives. The definition offered is that A is evolutionarily stable just in case both:

(i) U(A|A) ≥ U(B|A) (for all B different from A) and

(ii) If U(A|A) = U(B|A) then U(A|B)>U(B|B)

The leading idea of Maynard-Smith and Price only makes sense for mixed strategies if individuals play randomized strategies. But the replicator dynamics is appropriate for a model where individuals play pure strategies, and the counterparts of the mixed strategies of game theory are polymorphic states of the population. We retain this model, but introduce the notion of an evolutionarily stable state of the population. It is a polymorphic state of the population, which would satisfy the foregoing inequalities if the corresponding randomized strategies were used. An evolutionarily stable *strategy* corresponds to an evolutionarily stable *state* (ESS) in which the whole population uses that strategy.

Nash from Nature

Condition (i) in the definition of ESS looks a lot like the definition of Nash equilibrium. It is, in fact, the condition that <A, A> is a Nash equilibrium of the associated two person game in which both players have the payoffs specified in by the fitness matrix, U(_|_). the second condition adds a kind of stability requirement. The requirement is sufficient to guarantee strong dynamical stability in the replicator dynamics:

EVERY ESS is a STRONGLY DYNAMICALLY STABLE (or ATTRACTING) EQUILIBRIUM in the REPLICATOR DYNAMICS

This is more than sufficient to guarantee Nash equilibrium in the corresponding game:

> IF A is a DYNAMICALLY STABLE EQUILIBRIUM in the REPLICATOR DYNAMICS
> then <A,A> is a NASH EQUILIBRIUM of the corresponding TWO-PERSON GAME.

(The converse of each of the foregoing propositions fails. For more details see Hofbauer and Sigmund, 1988; van Damme, 1987; Weibull 1997).

Evidently, Nash equilibrium has an important role to play here, in the absence of common knowledge of rationality or even rationality itself. The reason is quite clear, but nevertheless deserves to be emphasized. The underlying dynamics is adaptive: it has a tendency towards maximal fitness. Many alternative dynamics — of learning as well as of evolution — share this property. (See Borgers and Sarin, 1997, for a connection between reinforcement learning and replicator dynamics, but compare the discussion of Fudenberg and Levine, 1998, Ch. 3). There is a moral here for philosophers and political theorists who have attacked the theory of games on the basis of its rationality assumptions. Game theory has a far broader domain of application than that suggested by its classical foundations.

Symmetry

For every evolutionary game, there is a corresponding symmetric two-person game, and for every ESS in the Evolutionary game, there is a corresponding symmetric Nash equilibrium in the two-person game. Symmetry is imposed on the Nash equilibrium of the two-person game because the players have no identity in the evolutionary game. Different individuals play the game. The things which have enduring identity are the strategies. Evolutionary games are played under what I call 'The Darwinian Veil of Ignorance' in Chapter I of my book, *Evolution of the Social Contract*. Evolution ignores individual idiosyncratic concerns simply because individuals do not persist through evolutionary time.

For example, consider the game of Chicken. There are two strategies: Swerve; Don't. The fitnesses are:

$$U(S|S) = 20$$
$$U(S|D) = 15$$
$$U(D|S) = 25$$
$$U(D|D) = 10$$

In the two-person game, there are two Nash equilibria in pure strategies: player one swerves and player two doesn't, player two swerves and player one doesn't. There is also a mixed Nash equilibrium with each player having equal chances of swerving. In the evolutionary setting, there are just swervers and non-swervers. The only equilibrium of the two-person game that corresponds to an ESS of the evolutionary game is the mixed strategy. It corresponds to an evolutionary stable polymorphic state where the population is equally split between swervers and non-swervers. Any departure from the state is rectified by the replicator dynamics, for it is better to swerve when the majority don't and better not to swerve when the majority do.

The evolutionary setting has radically changed the dynamical picture. If we were considering learning dynamics for two fixed individuals, the relevant state space would be the unit square with the x-axis representing the probability that player one would swerve and the y-axis representing the probabilities that player two would swerve. With any reasonable learning dynamics, the mixed equilibrium of the two-person game would be highly unstable and the two pure equilibria would be strongly stable. The move to the evolutionary setting in effect restricts the dynamics to the diagonal of the unit square. The probability that player one will encounter a given strategy must be the same as the probability that player two will. It is just the proportion of the population using that strategy. On the diagonal, the mixed equilibrium is now strongly stable.

For another example which is of considerable importance for social philosophy, consider the simplest version of the Nash bargaining game. Two individuals have a resource to divide. We assume that payoff just equals the amount of the resource. They each must independently state the minimal fraction of the resource that they will accept. If these amounts add up to more than the total resource, there is no bargain struck and each players gets nothing. Otherwise, each gets what she asks for.

This two person game has an infinite number of Nash equilibria of the form: Player one demands x of the resource and player two demands (1-x) of the resource ($0<x<1$). Each of these Nash equilibria is strict — which is to say that a unilateral deviation from the equilibrium not only produces no gain; it produces a positive loss. Here we have the problem of multiple Nash equilibria in especially difficult form. There are an infinite number of equilibria and, being strict, they satisfy all sorts of refinements of the Nash equilibrium concept (see van Damme, 1987).

Suppose we now put this game in the evolution context that we have developed. What pure strategies are evolutionarily stable? There is exactly one: Demand half! First, it is evolutionarily stable. In a population in which all demand half, all get half. A mutant who demanded more of the natives would get nothing; a mutant who demanded less would get less. Next, no other pure strategy is evolutionarily stable. Assume a population composed of players who demand x, where $x<1/2$. Mutants who demand 1/2 of the natives will get 1/2 and can invade. Next consider a population of players who demand x, where $x>1/2$. They get nothing. Mutants who demand y, where $0<y<(1-x)$ of the natives will get y and can invade. So can mutants who demand 1/2, for although they get nothing in encounters with natives, they get 1/2 in encounters with each other. Likewise, they can invade a population of natives who all demand 1. Here the symmetry requirement imposed by the evolutionary setting by itself selects a unique equilibrium from the infinite number of strict Nash equilibria of the two-person game. The 'Darwinian Veil of Ignorance' gives an egalitarian solution.

This is only the beginning of the story of the evolutionary dynamics of bargaining games. Even in the game I described, there are evolutionarily stable polymorphic states of the population which may be of considerable interest. (But see Alexander and Skyrms, 1999, and Alexander, 1999, for local interaction models where these polymorphisms almost never occur.) And in more complicated evolutionary games we can consider individuals who can occupy different roles, with the payoff function of the resource being role-dependent. However, at the most basic level, we have a powerful illustration of my point. The evolutionary setting for game theory here

makes a dramatic difference in equilibrium selection, even while it supports a selection of a Nash equilibrium.

Weakly Dominated Strategies

The strategic situation can be radically changed in bargaining situations by introducing sequential structure. Consider the Ultimatum game of Güth, Schmittberger and Schwarze (1982). One player — the Proposer — demands some fraction of the resource. The second player — the Responder — is informed of the Proposer's proposal and either takes it or leaves it. If he takes it the first player gets what she demanded and he gets the rest. If he leaves it, neither player gets anything.

The are again an infinite number of Nash equilibria in this game, but from the point of view of rational decision theory they are definitely not created equal. For example, there is a Nash equilibrium where the Proposer has the strategy 'Demand half' and the Responder has the strategy 'Accept half or more but reject less'. Given each player's strategy, the other could not do better by altering her strategy. But there is nevertheless something odd about the Responder's strategy. This can be brought out in various ways. In the first place, the Responder's strategy is weakly dominated. That is to say that there are alternative strategies which do as well against all possible Proposer's strategies, and better against some. For example, it is weakly dominated by the strategy 'Accept all offers'. A closely related point is that the equilibrium at issue is not subgame perfect in the sense of Selten (1965). If the Proposer were to demand 60 per cent, this would put the responder into a subgame, which in this case would be a simple decision problem: 40 per cent or nothing. The Nash equilibrium of the subgame is the optimal act: Accept 40 per cent. So the conjectured equilibrium induces play on the subgame which is not an equilibrium of the subgame.

For simplicity, let's modify the game. There are ten lumps of resource to divide; lumps are equally valuable and can't be split; the proposer can't demand all ten lumps. Now there is only one subgame perfect equilibrium — the Proposer demands nine lumps and the Responder has the strategy of accepting all offers. If there is to be a Bayesian rational response to any possible offer, the Responder must have the strategy of accepting all offers. And if the Proposer knows that the Responder will respond by optimizing no matter what she does, she will demand nine lumps.

It is worth noting that the subgame perfect equilibrium predicted by the foregoing rationality assumptions does not seem to be what is found in experimental trials of the ultimatum game — although the interpretation of the experimental evidence is a matter of considerable controversy in the literature. Proposers usually demand less than 9 and often demand five. Responders often reject low offers, in effect choosing zero over one or two. I do not want to discuss the interpretation of these results here. I only note their existence. What happens when we transpose the ultimatum game into an evolutionary setting?

Here the game itself is not symmetric — the proposer and responder have different strategy sets. There are two ways to fit this asymmetry into an evolutionary framework. One is to model the game as an interaction between two disjoint populations. A proposer is drawn at random from the proposer population and a responder is drawn at random from the responder population and they play the game with the payoff being expected number of offspring. The alternative is a single population model with

roles. Individuals from a single population sometimes are in the role of Proposer and sometimes in the role of Responder. In this model, individuals are required to have the more complex strategies appropriate to the symmetrized game, for example:

If in the role of Proposer demand x;
If in the role of Responder and proposer demands z, take it;
Else if in the role of Responder and proposer demands z′, take it;
Else if in the role of responder and the proposer demands z′′,leave it;
etc.

The evolutionary dynamics of the two population model was investigated by Binmore, Gale and Samuelson (1995) and that of the one population symmetrized model by myself (1996; 1998b) for the replicator flow and by Harms (1994; 1997) for the Darwin map.

Despite some differences in modelling, all these studies confirm one central fact. Evolutionary dynamics need not eliminate weakly dominated strategies; evolutionary dynamics need not lead to subgame perfect equilibrium. Let me describe my results for a small game of Divide Ten, where proposers are restricted to two possible strategies: Demand Nine; Demand Five. Responders now have four possible strategies depending on how they respond to a demand of nine and how they respond to a demand of five. The typical result of evolution starting from a population in which all strategies are represented is a polymorphic population which includes weakly dominated strategies.

One sort of polymorphism includes Fairmen types who demand five and accept five but reject greedy proposals; together with Easy Riders who demand five and accept all. The Fairman strategy is weakly dominated by the Easy Rider strategy, but nevertheless some proportion of Fairmen can persist in the final polymorphism. Another sort of polymorphism consists of Gamesmen who demand nine and accept all, together with Mad Dogs who accede to a demand of nine but reject a Fairman's demand of five. Mad Dog is weakly dominated by Gamesman but nevertheless some proportion of Mad Dogs can persist in the final polymorphism. Which polymorphism one ends up with depends on what population proportions one starts with. However, in either case one ends up with populations which include weakly dominated strategies.

How is this possible? If we start with a completely mixed population — in which all strategies are represented — the weakly dominated strategies must have a smaller average fitness than the strategies which weakly dominate them. The weakly dominating strategies by definition do at least as well against all opponents and better against some. Call the latter the Discriminating Opponents. As long as the Discriminating Opponents are present in the population, the weakly dominating do better than the weakly dominated, but the Discriminating Opponents may go extinct more rapidly than the weakly dominated ones. This leaves a polymorphism of types which do equally well in the absence of Discriminating Opponents. This theoretical possibility is, in fact, the typical case in the ultimatum game. This conclusion is not changed, but rather reinforced, if we enrich our model by permitting the Proposer to demand 1, 2, 3, 4, 5, 6, 7, 8, 9. Then we get more complicated polymorphisms which typically include a number of weakly dominated strategies.

It might be natural to expect that adding a little mutation to the model would get rid of the weakly dominated strategies. The surprising fact is that such is not the case.

The persistence of weakly dominated strategies here is quite robust to mutation. It is true that adding a little mutation may keep the population completely mixed, so that weakly dominated strategies get a strictly smaller average fitness than those strategies which weakly dominate them, although the differential may be very small. But mutation also has a dynamical effect. Other strategies mutate into the weakly dominated ones. This effect is also very small. In polymorphic equilibria under mutation these two very small effects come into balance.

That is not to say that the effects of mutation are negligible. These effects depend on the mutation matrix, Mij, of probabilities that one given type will mutate into another given type. We model the combined effects of differential reproduction and mutation by a modification of the Darwin Flow:

$$dP(A_i)/dt = (1-e)[P(A_i)\,(U(A_i)-UBAR)/UBAR] + e[SUM_j\, P(A_j)\, M_{ij} - P(A_i)]$$

(See Hofbauer and Sigmund, 1988, p. 252)

There is no principled reason why these probabilities should all be equal. But let us take the case of a uniform mutation matrix as an example. Then, in the ultimatum game, introduction of mutation can undermine a polymorphism of Fairmen and Easy Riders and lead to a polymorphism dominated by Gamesmen and Mad Dogs. With a uniform mutation matrix and a mutation rate of 0.001, there is a polymorphism of about 80 per cent gamesmen and 20 per cent Mad Dogs — with other types maintained at very small levels by mutation. Notice that the weakly dominated strategy Mad Dog persists here at quite substantial levels. Other mutation matrices can support polymorphisms consisting mainly of Fairmen and Free Riders. In either case, we have equilibria that involve weakly dominated strategies.

A crack has appeared between game theory based on rationality and game theory based on evolutionary dynamics. There are rationality-based arguments against weakly dominated strategies and for sub-game perfect equilibrium. We have seen that evolutionary dynamics does not respect these arguments. This is true for both one-population and two-population models. It is true for both continuous and discrete versions of the dynamics. It remains true if mutation is added to the dynamics. I will not go into the matter here, but it also remains true if variation due to recombination — as in the genetic algorithm — is added to the dynamics (see Skyrms, 1996, Ch. 2). We shall see in the next section that the crack widens into a gulf if we relax the assumption of random pairing from the population.

Strongly Dominated Strategies

Consider a two-person game in which we assume that the players are Bayesian rational and know the structure of the game, but nothing more. We do not assume that the players know each other's strategies. We do not assume that Bayesian rationality is Common Knowledge. In general, this assumption will not suffice for Nash equilibrium or even Rationalizability. But one thing that it does is guarantee that players will not play strongly dominated strategies.

In certain special games, this is enough to single out a unique Nash equilibrium. The most well known game of this class is the Prisoner's Dilemma. In this game both players must choose between co-operation [C] and defection [D]. For each player:

$$U(C|C) = 10$$
$$U(D|C) = 15$$
$$U(C|D) = 0$$
$$U(D|D) = 5$$

Defection strongly dominates co-operation and if players optimize they both defect.

If we simply transpose our game to an evolutionary setting, nothing is changed. The game is symmetric; players have the same strategy sets and the payoff for one strategy against another does not depend on which player plays which. Evolutionary dynamics drives the co-operators to extinction. Defection is the unique evolutionarily stable strategy and a population composed exclusively of defectors is the unique evolutionarily stable state. So far evolution and rational decision theory agree.

Let us recall, however, that our evolutionary model was based on a number of simplifying assumptions. Among those was the assumption that individuals are paired *at random* from the population to play the game. This may or may not be plausible. There is a rich biological literature discussing situations where it is not plausible for biological evolution (see Hamilton, 1964; Sober and Wilson, 1998). With regard to cultural evolution I believe that many social institutions exist in order to facilitate non-random pairing (Milgrom *et al*., 1990). A more general evolutionary game theory will allow for correlated pairing (Skyrms, 1996, Ch. 3).

Here the fundamental objects are conditional pairing probabilities, P(A|B), specifying the probability that someone meets an A-player given that she is a B-player. These conditional pairing proportions may depend on various factors, depending on the particular biological or social context being modelled. The expected fitness of a strategy is now calculated using these conditional probabilities:

$$U(A) = SUM_i\ U(A|B_i)\ P(B_i|A)$$

Now suppose that nature has somehow — I don't care how — arranged high correlation between like strategies among individuals playing the Prisoner's Dilemma. For instance, suppose:

$$P(C|C) = P(D|D) = 0.9 \text{ and } P(C|D) = P(D|C) = 0.1$$

Then the fitness of co-operation exceeds that of defection and the strategy of co-operation takes over the population. Correlated evolutionary game theory permits models in which a strongly dominated strategy is selected.

We can, of course, consider correlation in two person games between rational decision makers. This was done by Aumann (1974; 1987) in his seminal work on correlated equilibrium. Aumann takes the natural step of letting the 'coin flips' of players playing mixed strategies be correlated. The resulting profile is a joint correlated strategy. Players know the joint probability distribution and find out the results of their own 'coin flips'. If, whatever the results of those flips, they do not gain in expected utility by unilateral deviation, they are at a correlated equilibrium. However, this natural generalization is quite different than the generalization of mixed strategies as polymorphic populations that I have sketched. In particular, there is only one Aumann correlated equilibrium in Prisoner's Dilemma. It has each player defecting. More generally, Aumann correlated equilibria do not use strongly dominated strategies.

In fact, evolutionary game theory can deal with two kinds of mixed strategy. The first kind arises when individuals themselves use randomized strategies. The second kind interprets a population polymorphism as a mixed strategy. I have focused on the second kind in this paper. We have assumed that individuals play pure strategies in order to preserve the rationale for the replicator dynamics. If one drops the independence assumption from the first kind of mixed strategy, one gets Aumann correlated equilibrium. If one drops the independence assumption from the second kind, one gets the kind of correlated evolutionary game theory I am discussing here. In this setting, new phenomena are possible — the most dramatic of which include the fixation of strongly dominated strategies.

Prisoner's Dilemma is so widely discussed because it is a simple paradigm of the possibility of conflict between efficiency and strict dominance. Everyone would be better off if everyone co-operated. Co-operation is the efficient strategy. Whatever the other player does, you are better off defecting. Defection is the dominant strategy. In the game theory of rational choice, dominance wins. But in correlated evolutionary game theory, under favourable conditions of correlation, efficiency can win.

The point is general. If there is a strategy such that it is best if everyone uses it, then sufficiently high auto-correlation will favour that strategy. Perfect correlation imposes a kind of Darwinian categorical imperative under these conditions. Others do unto you as you do unto them. Then a strategy, A, such that $U(A|A)>U(B|B)$ for all B different from A, will be carried to fixation by the replicator dynamics from any initial state in which it is represented in the population.

On the other hand, in some strategic situations, anti-correlation may promote social welfare. Suppose that there are two types of a species that do not do well when interacting with themselves, but each of which gains a large payoff when interacting with the other type. Then the payoff to each type and the average payoff to the species might be maximized if the types could somehow anti-correlate. The situation hypothesized is one where mechanisms for detection might well evolve to support pairing with the other type. Consider species that reproduce sexually.

The introduction of correlation, in the manner indicated, takes evolutionary game theory into largely uncharted waters — unexplored by the traditional game theory based on rational choice. The basic structure of the theory is relaxed, opening up novel possibilities for the achievement of efficiency. One of the simplest and most dramatic examples is the possibility of the evolution of the strongly dominated strategy of co-operation in the one-shot Prisoner's Dilemma. (For various types of correlation generated by learning dynamics see Vanderschraaf and Skyrms, 1994. For correlation generated by spatial interaction, see Alexander, 1999, and Alexander and Skyrms, 1999. For the role of correlation in a general treatment of convention see Vanderschraaf, 1995; 1998.)

Utility and Rationality

From the fundamental insight that, in the random pairing model Darwin supports Nash, we have moved to an appreciation of ways in which game theory based on rational choice and game theory based on evolution may disagree. In the Ultimatum game, evolution does not respect weak dominance or subgame perfection. In the Prisoner's Dilemma with correlation evolution may not respect strong dominance.

Is this an argument for the evolution of irrationality? It would be a mistake to leap to this conclusion. Irrationality in Bayesian terms does not consist in failing to maximize Darwinian fitness, but rather in failing to maximize expected utility. One can conjecture utility functions which save the Bayes rationality of *prima facie* irrational behaviour in putative Ultimatum or Prisoner's Dilemma games. Whether this is the best way to explain observed behaviour remains to be seen. But if it were, we could talk about the evolution of utility-functions that disagree with Darwinian fitness rather than the evolution of irrationality.

However that may be, when we look at the *structure* of game theory based on rational choice and that of game theory based on evolutionary dynamics, we find that beyond the areas of agreement there are also areas of radical difference.

Avoiding the Curse of Symmetry

Let me close by returning to the example with which I started. That is the coordination game where players get a positive payoff if, and only if, they choose the same colour. Suppose it is a population that is evolving a custom and the possibilities are 'Choose red' and 'Choose green'. Corresponding to the mixed strategy delivered up by hyper-rational equilibrium selection, there is a population state where half the population chooses red and half chooses green. If we assume random encounters and the replicator dynamics, this population state is indeed a dynamic equilibrium, but it is dynamically unstable. A population in this state is like a ball rolling along a knife-edge. If the population has a little greater proportion on the red side, the replicator dynamics will carry it to a state where all choose red; if it has a little greater proportion on the green side, the dynamics will carry it to a state where all choose green. Variant adaptive dynamics will do the same.

This is a simplified picture, which encapsulates essential features of the evolution of conventions in general. One case of special interest is the evolution of the meanings of symbols in the kind of sender-receiver games introduced by David Lewis (1969). Rational-choice equilibrium selection theory would have to lead players to randomized 'babbling equilibria' where no meaning at all is generated. Evolutionary dynamics leads to the fixation of meaning in signalling system equilibria. You can read about it in Skyrms (1996, Ch.5; 1999).

Conclusion

As an explanatory theory of human behaviour, dynamical models of cultural evolution and social learning hold more promise of success than models based on rational choice. Under the right conditions, evolutionary models supply a rationale for Nash equilibrium that rational choice theory is hard-pressed to deliver. Furthermore, in cases with multiple symmetrical Nash equilibria, the dynamic models offer a plausible, historically path-dependent model of equilibrium selection. In conditions, such as those of correlated encounters, where the evolutionary dynamic theory is structurally at odds with the rational choice theory, the evolutionary theory provides the best account of human behaviour.

References

Alexander, J. (1999), 'The (spatial) evolution of the equal split', Working Paper Institute for Mathematical Behavioural Sciences U.C.Irvine.

Alexander, J. and Skyrms, B. (1999), 'Bargaining with neighbors: is justice contagious?', Working Paper Logic and Philosophy of Science U.C.Irvine.

Aumann, R. J. (1974), 'Subjectivity and correlation in randomized strategies', *Journal of Mathematical Economics*, **1**, pp. 67–96.

Aumann, R. J. (1987), 'Correlated equilibrium as an expression of Bayesian rationality', *Econometrica* **55**, pp. 1–18.

Binmore, K. (1993), 'Game theory and the social contract Vol. 1', *Playing Fair* (Cambridge, MA: MIT Press).

Binmore, K. (1998), 'Game theory and the social contract Vol. 2', *Just Playing* (Cambridge,MA: MIT Press).

Binmore, K., Gale, J. and Samuelson, L. (1995), 'Learning to be imperfect:The ultimatum game', *Games and Economic Behaviour*, **8**, pp. 56–90.

Bernheim, B. D. (1984), 'Rationalizable strategic behaviour', *Econometrica* **52**, pp. 1007–28.

Bjornerstedt, J. and Weibull, J. (1995), 'Nash equilibrium and evolution by imitation', in *The Rational Foundations of Economic Behavior*, ed. K. Arrow *et al.* (New York: MacMillan), pp. 155–71.

Bomze, I. (1986), 'Non-cooperative two-person games in biology: A classification', *International Journal of Game Theory*, **15**, pp. 31–57.

Borgers, T. and Sarin, R. (1997), 'Learning through reinforcement and the replicator dynamics', *Journal of Economic Theory*, **77**, pp. 1–14.

Camerer, V. (1997), 'Progress in behavioural game theory', *Journal of Economic Perspectives*, **11**, pp. 167–88.

Duffy, J. and Nagel, R. (1997), 'On the robustness of behaviour in experimental "Beauty Contest" games', *The Economic Journal*, **107**, pp. 1684–700.

Fudenberg, D. and Levine, D. (1998), *The Theory of Learning in Games* (MIT :Cambridge, MA).

Gauthier, D. (1969), *The Logic of the Leviathan* (Oxford: Oxford University Press).

Gauthier, D. (1986), *Morals by Agreement* (Oxford: Clarendon Press).

Gibbard, A. (1990), *Wise Choices, Apt Feelings:A Theory of Normative Judgement* (Oxford: Clarendon Press).

Güth, W., Schmittberger, R. and Schwarze, B. (1982), 'An experimental analysis of ultimatum bargainin',g *Journal of Economic Behaviour and Organization*, **3**, pp. 367–88.

Güth, W., and Tietz, R. (1990), 'Ultimatum bargaining behaviour: A survey and comparison of experimental results', *Journal of Economic Psychology*, **11**, pp. 417–49.

Hamilton, W. D. (1964), 'The genetical evolution of social behaviour', *Journal of Theoretical Biology*, **7**, pp. 1–52.

Harms, W. (1994), 'Discrete replicator dynamics for the ultimatum game with mutation and recombination', Technical Report, (University of California, Irvine).

Harms, W. (1997), 'Evolution and ultimatum bargaining', *Theory and Decision*, **42**, pp. 147–75.

Harsanyi, J. and Selten, R. (1988) *A General Theory of Equilibrium Selection in Games* (Cambridge, MA: MIT Press).

Ho, T. H., Weigelt, K. and Camerer, C. (1996), 'Iterated dominance and iterated best-response in experimental *p*-beauty contests', Social Science Working Paper 974, California Institute of Technology.

Hofbauer, J. and Sigmund, K. (1988), *The Theory of Evolution and Dynamical Systems* (Cambridge: Cambridge University Press).

Hume, D. (1975), *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, reprinted from the posthumous edition of 1777 with text revised and notes by P. H. Nidditch (Oxford:Clarendon Press).

Keynes, J.M. (1936), *The General Theory of Employment, Interest and Money* (New York: Harcourt Brace).

Lewis, D. (1969), *Convention* (Cambridge, MA:Harvard University Press).

Maynard-Smith, J. and Price, G.R. (1973), 'The logic of animal conflict', *Nature*, **146**, pp. 15–18.

Maynard-Smith, J. and Parker, G.R. (1976), 'The logic of asymmetric contests', *Animal Behaviour*, **24**, pp. 159–75.

Milgrom, P., North, D. and Weingast, B. (1990), 'The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs', *Economics and Politics*, **2**, pp. 1–23.

Moulin, H. (1986), *Game Theory for the Social Sciences* (New York: New York University Press).
Nagel, R. (1995), 'Unravelling in guessing games:An experimental study', *American Economic Review*, **85**, pp. 1313–26.
Pearce, D.G. (1984), 'Rationalizable strategic behaviour and the problem of perfectio',n *Econometrica*, **52**, pp. 1029–50.
Sacco, P.L. (1995), 'Comment', in *The Rational Foundations of Economic Behavior*, ed. K. Arrow *et al.* (New York: Macmillan).
Samuelson, L. (1997), *Evolutionary Games and Equilibrium Selection* (Cambridge,MA:MIT Press).
Samuelson, L. (1988), 'Evolutionary foundations of solution concepts for finite two-player normal form games', in *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, ed. M. Vardi (Los Altos, CA: Morgan Kaufmann).
Schlag, K. (1994), 'Why imitate and if so how? Exploring a model of social evolution', *Discussion Paper B-296* (Department of Economics, University of Bonn).
Schlag, K. (1996), 'Why imitate and if so, how? A bounded rational approach to many armed bandits', *Discussion Paper B-361* (Department of Economics, University of Bonn).
Selten, R. (1975), 'Re-examination of the perfectness concept of equilibrium in extensive games', *International Journal of Game Theory*, **4**, pp. 25–55.
Selten, R. (1965), 'Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit', *Zeitschrift fur die gesamte Staatswissenschaft*, **121**, pp. 301–24, 667–89.
Skyrms, B. (1990), *The Dynamics of Rational Deliberation* (Cambridge,MA: Harvard University Press).
Skyrms, B. (1994a), 'Darwin meets 'The logic of decision': Correlation in evolutionary game theory', *Philosophy of Science*, **61**, pp. 503–28.
Skyrms, B. (1994b), 'Sex and justice', *The Journal of Philosophy*, **91**, pp. 305–20.
Skyrms, B. (1995), 'Introduction to the Nobel symposium on game theory', *Games and Economic Behaviour*, **8**, pp. 3–5.
Skyrms, B. (1996), *Evolution of the Social Contract* (New York: Cambridge University Press).
Skyrms, B. (1998a), 'Mutual aid' in *Modelling Rationality, Morality and Evolution*, ed. Peter Danielson (Oxford: Oxford University Press).
Skyrms, B. (1998b), 'Evolution of an anomaly', *Protosoziologie*, **12**, pp. 192–211.
Skyrms, B. (1999), 'Evolution of inference', in *Dynamics in Human and Primate Societies*, ed. T. Kohler and G. Gumerman. SFI Studies in the Sciences of Complexity. (New York: Oxford University Press).
Sober, E. (1992), 'The evolution of altruism: Correlation, cost and benefit', *Biology and Philosophy*, **7**, pp. 177–87.
Sober, E. and Wilson, D.S. (1998), *Unto Others:The Evolution and Psychology of Unselfish behaviour* (Cambridge, MA: Harvard University Press).
Spinoza, B. (1985), *Ethics:The Collected Works of Spinoza, Volume I*, trans. E. Curley (Princeton: Princeton University Press).
Stahl, D.O. (1996), 'Boundedly rational rule learning in a guessing game', *Games and Economic behaviour*, **16**, pp. 303–30.
Sugden, R. (1986), *The Economics of Rights, Co-operation and Welfare* (Oxford: Blackwell).
Taylor, P. and Jonker, L. (1978), 'Evolutionarily stable strategies and game dynamics', *Mathematical Biosciences*, **16**, pp.76–83.
van Damme, E. (1987), *Stability and Perfection of Nash Equilibria* (Berlin: Springer).
Vanderschraaf, P. (1998), 'Knowledge, equilibrium and convention', *Erkenntnis*, **49**, pp. 337–69.
Vanderschraaf, P. (1995), 'Convention as correlated equilibrium', *Erkenntnis*, **42**, pp. 65–87.
Vanderschraaf, P. and Skyrms, B. (1994), 'Deliberational correlated equilibrium', *Philosophical Topics*, **21**, pp. 191–227.
von Neumann, J. and Morgenstern, O. (1947), *Theory of Games and Economic Behaviour, 2nd. ed.* (Princeton: Princeton University Press).
Weibull, J. (1997), *Evolutionary Game Theory* (Cambridge, MA: MIT Press).