Game Theory, Evolution and Justice

Peter Vanderschraaf

December 1998

Technical Report No. CMU-PHIL-95

Philosophy
Methodology
Logic

Carnegie Mellon

Pittsburgh, Pennsylvania 15213

Game Theory, Evolution and Justice[1]

Peter Vanderschraaf

§0. Introduction

In 1954, Richard Braithwaite delivered a lecture at Cambridge University entitled *Theory of Games as a Tool for the Moral Philosopher.*[2] Braithwaite had an idea for developing an account of distributive justice which could resolve practical moral problems. He argued that problems of distributive justice have the structure of the *bargaining problem* recently analyzed by John Nash (1950*b*, 1953). Only a decade prior to Braithwaite's Cambridge lecture, John von Neumann and Oskar Morgenstern (1944) had established an important new branch of social science with the publication of their treatise *Theory of Games and Economic Behavior.* Von Neumann and Morgenstern laid the foundation for a scientific theory of social interaction, which they conjectured might one day be as rigorous and predictively powerful as the mathematical physics of their time. In the early 1950s, Nash made tremendous strides in this research program, including his formal theory of bargaining. Game theory has developed far beyond the classical theory of von Neumann, Morgenstern and Nash, and now spans many disciplines. Remarkably, quite early in its development Braithwaite recognized the significance of game theory for moral philosophy, and used elements of Nash's bargaining theory to derive his principle of distributive justice.

In his closing remarks, Braithwaite speculated that game theory might in time transform moral philosophy, much as statistics had transformed the social sciences. Such a transformation has yet to occur --- Braithwaite thought it might take centuries --- but

[1] Thanks to Andrew Artz, Sharon Lloyd, Kay Mathiesen and Robert Sugden for their many valuable comments on early versions of this essay.

[2] This was Braithwaite's inaugural lecture following his election to the Knightbridge Professor of Moral Philosophy at Cambridge.

game-theoretic arguments have made important inroads into contemporary ethics and political philosophy. Much progress has been made in this general research program by linking game theory with concepts from evolutionary theory. In the late 1970s and 1980s J. L. Mackie (1978), Robert Axelrod (1981, 1984) and Robert Sugden (1986) argued that the *evolutionary game theory* introduced by the biologists John Maynard Smith and G. R. Price (Maynard Smith and Price 1973, Maynard Smith 1982) to model behavioral patterns in nonhuman species could be adapted to serve the moral philosopher.[3] Axelrod and Sugden characterized social norms as *equilibria* of games, which gradually emerge as the result of some evolutionary process. In the years following Axelrod's and Sugden's pioneering works, the interplay between evolutionary game theory and moral philosophy has increased, and evolutionary game theory itself has enjoyed a period of explosive growth. Recently two landmark works have appeared which present the elements of a theory of the social contract based upon evolutionary game theory: Brian Skyrms' *Evolution of the Social Contract* (1996) and Ken Binmore's *Just Playing* (1998), the sequel to his *Playing Fair* (1994). These works reflect their authors' unusual combination of intellectual gifts: a deep knowledge of both game theory and moral philosophy, an uncanny ability to glean out the central unifying issues in a mass of difficult relevant literature, and startling creativity. They draw upon the cumulation of research into game theory and its uses in ethics following Braithwaite's Cambridge lecture, and are also major contributions to both moral philosophy and to evolutionary game theory. In this paper, I will review some of the central issues dealt with in Skyrms' and Binmore's works, which represent the state of the art of game-theoretic reasoning in moral philosophy.

Game theory is a formal logic of interactive decisions. Given how one interprets this logic, one can use it either to explain why certain patterns of behavior prevail in a

[3] Mackie (1978, p. 453) was the first to suggest that the study of certain "evolutionary games" might prove illuminating for moral philosophers, without referring to Maynard Smith and Price's 1973 article which introduced evolutionary game theory. Axelrod and Sugden arrived at the idea of analyzing social norms in terms of evolutionary game theory independently of Mackie's article.

community or to argue for a precise account of how individuals ought to behave in various contexts, as Braithwaite did for the case of sharing scarce resources. Skyrms uses evolutionary game theory to account for how many social institutions, including norms of fairness and reciprocity, could have evolved. Binmore presents a general argument based upon evolutionary game theory for how a society should progress from its current social contract. Binmore's normative recommendations are a powerful alternative to social contract theories like those of Rawls (1957, 1971), Harsanyi (1955, 1977) and Gauthier (1986), which view the social contract as the object of a single and irrevocable choice under certain constraints. As we shall see below, the evolutionary perspective Skyrms and Binmore adopt avoids many of the pitfalls of the binding-choice approach to the social contract. Game theory also illuminates the relationship between morality and rational choice. A game is "solved" when the actions of the agents engaged in the game form an equilibrium, that is, each agent's choice of action is rational given the others' actions. To the extent that philosophers can show that conformity with moral norms constitutes an equilibrium of a corresponding game, they have a solution to the old problem of reconciling morality and self-interest. As Binmore and Skyrms both observe, this program builds upon David Hume's idea of explaining justice as a system of conventions. By basing their arguments upon evolutionary game theory, Skyrms and Binmore can account for why individuals would willingly fulfill the requirements of a social contract, which is the crucial point on which binding-choice social contract theories like Rawls' falter. On the other hand, in many important social settings game theory apparently shows that following the requirements of morality contradicts equilibrium behavior. This is not, as some have claimed, a defect of game theory. Morality and self-interest might not coincide perfectly, and game theory helps the moral philosopher better understand the limits of this coincidence. Skyrms' and Binmore's works raise the analysis of how well rationality meshes with morality to a new level of sophistication.

Philosophers following Braithwaite's footsteps have always faced a serious obstacle, namely, the *equilibrium selection problem.* If one identifies a convention or norm with an equilibrium of a particular game, this in no way explains the *origins* of norms. A game typically has a plurality of distinct equilibrium points. *How do individuals identify a given equilibrium as the one which defines the norm they ought to follow?* Sometimes one can argue that agents should follow a particular equilibrium on moral grounds. For instance, one might recommend a certain equilibrium because it allots each individual a fair share, or because it protects the disadvantaged. But in many cases, such moral criteria do not select a unique equilibrium. And one might even ask why we count criteria like fairness as normatively relevant to problems which we can model using game theory. As we shall see in the following sections, Skyrms and Binmore apply evolutionary game theory to analyze two special equilibrium selection problems, namely: (*i*) to explain how the social contract we know could have emerged, and (*ii*) to justify proposals for reforming the current social contract.

§1. Fair Division

> All men think justice to be a sort of equality; . . . For they say that what is just is just for someone and that it should be equal for equals. But there still remains a question: equality or inequality of what? Here is a difficulty which calls for philosophical speculation. (Aristotle, *Politics* 1282b18-22)

Two individuals, Matthew and Luke, are vying for shares of a limited resource. Each would like as much of the resource for his own use as possible, yet each would rather have the other receive a share of the resource than have it go to waste. And they do not necessarily place the same value upon the various ways in which they might share, or fail to share, the resource. Braithwaite used this motivating example to illustrate the game-theoretic approach to distributive justice.[4] What would be the just way to divide the resource?

Matthew and Luke can either take their case to an arbiter, or try to solve the problem themselves. Let us consider the latter approach first. Suppose for simplicity's sake that each has two available courses of action: *hawk* (H), which is to claim all of the resource, and *dove* (D), which is to claim none. Their situation is summarized by the matrix of Figure 1.

Figure 1. Resource Acquisition Problem

$$\begin{array}{cc|c|c|} & & \multicolumn{2}{c}{\text{Matthew}} \\ & & D & H \\ \hline \text{Luke} & D & (0,0) & (1,4) \\ \hline & H & (4,1) & (0,0) \\ \hline \end{array}$$

$D =$ modest, $H =$ greedy

This matrix characterizes a *game in strategic form*, in which D and H are Luke and Matthew's *pure strategies*.[5] The first (second) coordinate of each ordered pair or *payoff*

[4] Braithwaite (1955, pp. 8-9) presents his example with the following story: Matthew and Luke live in adjacent apartments. Luke plays the trumpet, while Matthew plays the piano. Unfortunately, the walls in their apartments are thin. Each can hear the other's playing almost as well as he can hear his own playing. Admittedly, each finds the other's playing somewhat pleasant, so that he would rather keep quiet and listen to the other play than endure the cacophony if they both play or the complete silence if neither plays. But what each wants most is to play his own instrument undisturbed. So to coordinate their acts, one must keep quiet while the other plays. In Braithwaite's story, the resource Matthew and Luke would like to somehow share is a fixed period of playing time.

[5] Strategic form games model interactions in which the agents choose their strategies without being able to causally influence or even observe each others' strategies. Given this lack of information regarding the others' choices, each agent must choose as if all were choosing independently and simultaneously. *Extensive form games* model interactions in which the agents move in sequence, and might be able to observe some of each other's moves. In a game of *perfect information*, every agent moves with full knowledge of all of the moves that have occurred prior to her move.

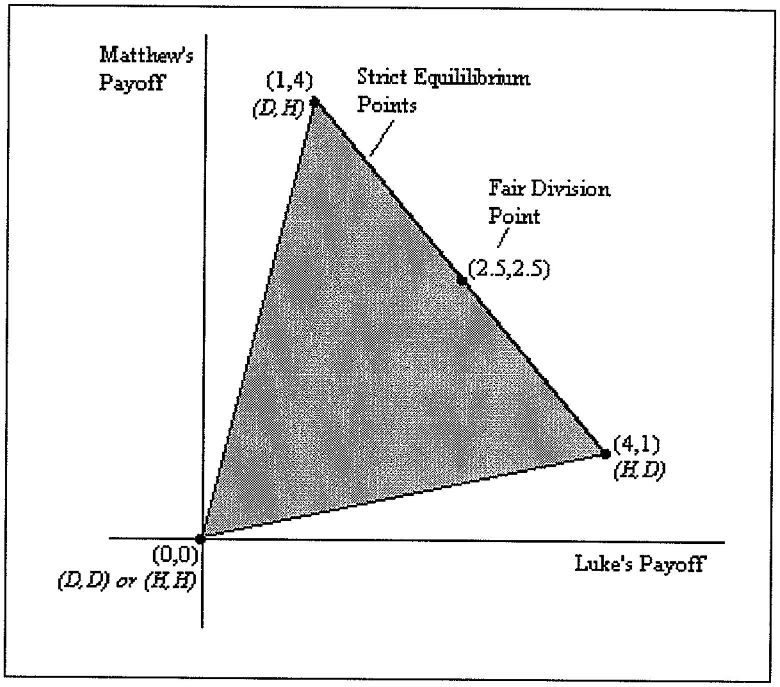
vector in the Figure 1 matrix is the row (column) agent's payoff. The payoffs reflect the desirability of the various possible outcomes for the agents.[6] Each agent is (Bayesian) rational if he chooses a strategy that yields his maximum expected payoff given his beliefs regarding his situation. In this game, each agent is rational to follow (D, H) if he expects the other to do likewise. Each agent is also rational to follow (H, D) if he expects the other to also follow (H, D), that is, (D, H) and (H, D) are *Nash equilibria* (1950a, 1951) in pure strategies of this game.[7] The Figure 1 game specifies an *impure* or *conflictual coordination problem*, since Matthew and Luke would like to coordinate on either (D, H) or (H, D), but Luke strictly prefers (H, D) while Matthew strictly prefers (D, H). So we seem to have no way to predict which equilibrium Matthew and Luke will follow.

Suppose next that Matthew and Luke's strategy options are enriched so that each can claim a portion of the resource. Luke's claim is compatible with Matthew's claim when their combined claims do not exceed the available amount of the resource. Each gets what he claims so long as their claims are compatible, and otherwise both get nothing. Their situation, which is an instance of Nash's bargaining problem, is depicted in Figure 2.

[6] Von Neumann and Morgenstern (1944) proved a representation theorem that showed that cardinal utilities can be derived for agents who follow certain axioms of decision making, including having preferences over *gambles* for goods, and who have a probability measure over the various outcomes of their decisions. *Von Neumann-Morgenstern utilities* are unique up to a choice of scale. Von Neumann and Morgenstern had partly rediscovered some of the results in Frank Ramsey's (1926) great essay 'Truth and Probability'. Ramsey proved a more general representation theorem which showed that if an agent has a sufficiently rich ordering of preferences over gambles and follows certain axioms of coherent decision making, then the agent can derive both a cardinal utility function *and* a subjective probability measure over the various outcomes of his decisions. L. J. Savage (1954) revived Ramsey's program, and proved a representation theorem that most economists now refer to when they ascribe subjective probabilities and cardinal utilities to rational agents.

[7] If Matthew and Luke can choose *mixed* strategies, that is, peg their pure strategies on probabilistically independent random devices, then this game has a third Nash equilibrium in mixed strategies in which Luke and Matthew each choose H with probability $\frac{4}{5}$. Unlike the pure strategy Nash equilibria, neither strategy of the mixed Nash equilibrium is a *unique* best reply to the other end of the mixed equilibrium. Nash (1950a, 1951) showed that every game with finitely many pure strategies has at least one Nash equilibrium if mixed strategies are allowed.

Figure 2. Fair Division Problem With Symmetric Payoffs



The shaded region is the set of payoff vectors corresponding to compatible claims by Matthew and Luke. To resolve their bargaining problem, we would expect each to claim half the resource. This agrees with the judgment they would expect from an arbiter. Aristotle observes that everyone agrees upon the *fair division principle*: A good should be distributed in proportion to relevant differences, and in particular equals should receive equal shares. Traditionally, philosophers have accepted the fair division principle without question, regarding disputes over the relevant criteria of difference as the "difficulty for philosophical speculation".

In *Evolution of the Social Contract*, Skyrms takes the fair division principle itself to be a matter for philosophical speculation. For this principle is not a consequence of the logic of rational choice. If Matthew and Luke each claim half the resource, then each receives half. If either of them were to deviate by claiming less, he would get less than half, while if he were to deviate by claiming more he would get nothing. So each agent is strictly better off to claim half given that the other claims half, that is, both claiming half is a *strict* Nash equilibrium. But if Matthew claims two thirds and Luke claims only one third, then they are at another strict Nash equilibrium, for Luke's claim is his unique best response to Matthew's claim and vice versa. Similarly, every pair of compatible claims that divides all of the resource is a strict Nash equilibrium. We might expect rational agents placed in fair division problems like the Figure 2 game to offer a multitude of different claims, depending on how cautious or aggressive they are. It is by no means a foregone conclusion that rational agents will follow the equilibrium corresponding to equal division. In a slightly different version of the Figure 2 game, fair division apparently should not occur at all. If Luke gets to propose a division of the resource first, and Matthew has only one chance to either accept or reject Luke's proposal, then for Luke to propose a fair division and for Matthew to accept is a Nash equilibrium. But Luke is now in the driver's seat. If Luke claims *all* the resource, then Matthew's unique best response is to accept, so Luke, knowing that Matthew is rational, should claim all and expect to receive all. This is the strict Nash equilibrium most favorable to Luke and least favorable to Matthew.[8]

As noted before, we tend to think of only one equilibrium in games like these as the *just* equilibrium, namely, the equilibrium of equal division. Our intuitions are confirmed both in everyday life and in the laboratory. In experimental studies, subjects tend to follow the Aristotelian norm of fair division in bargaining problems similar to the Figure 2 game. In experiments with a sequential structure that has one subject propose a division of a resource and the other "take it or leave it", to reject *any* proposal which leaves one with some positive share is to turn down something in favor of nothing. Still, proposers in such *ultimatum games* tend to offer a substantial portion of the resource to the other and frequently propose equal division, while those on the receiving end tend to

[8] This sequential game is of course a game of perfect information, because the second agent knows the first agent's proposal before responding.

reject proposals leaving them “too little”.[9] One can plausibly conclude from the experimental findings that the subjects are taking into account certain *norms of fairness* which guide their everyday dealings with others. But this brings back the question at hand. Why do we have fairness norms?

Skyrms suggests that we look to evolution to explain what rational choice cannot. Suppose that Matthew and Luke's encounter is but one of a series of similar encounters between members of a population who are paired at random in a given *base game* like the resource division game of Figure 2. Each time a pair enter into the base game, each agent employs a fixed strategy which she has “inherited” from others who have played the game previously. In biological evolution, members of a species inherit strategies genetically. In *cultural* evolution, the story is more complex, but the “inheritance” may stem in large part from what one has *learned* from observing others. Payoffs now reflect *reproductive fitness*. In biological evolution, high reproductive fitness is sometimes interpreted as a tendency to have many offspring, perhaps as the result of having a relatively long reproductive life. In cultural evolution, “reproductive” fitness can reflect a variety of accounts of how an individual might flourish, so that the more she tends to flourish the more others who follow her will tend to adopt the strategies she employs. Skyrms and Binmore do not argue for any particular account of the human good, but the payoffs in

[9] Roth (1995) surveys most of the known results in bargaining experiments. In an early experimental study of strategic form bargaining experiments, Nydegger and Owen (1975) had 30 pairs of subjects each play one of three games in which a pair had to agree upon a division of a sum of money, with the proviso that nonagreement would leave each with nothing. All 30 pairs agreed upon an equal division of the sum in question. Guth, Schmittberger and Schwarze (1982) were the first to publish results of an experiment with the ultimatum game. In their experiment, 42 subjects were divided into pairs, in which one played the role of the *allocator*, who was to propose a division of a sum of money, and the other played the role of the *recipient*, who could either accept or reject the division. A week later, the same subjects were invited to play the game again. In the first round, the modal proposal (made by 7 allocators) was 50% to both allocator and recipient, that is, equal division, and the mean of the proposals was 63% to the allocators and 37% to the recipients. In the second round, only two allocators proposed equal division, but the mean of the proposals was 32% to the recipients. Guth, Schmittberger and Schwarze's study was followed by a number of subsequent experimental studies which varied the conditions of the bargaining games to be played in order test the robustness of their results. In one of the most ambitious of these experimental studies, Ochs and Roth (1989) found that subjects in ultimatum games seldom follow the predictions of rational choice game theory, even when they play ultimatum games repeatedly.

evolutionary game theory are in principle a way to make precise the varying degrees to which an individual might achieve an ultimate good such as Aristotelian *eudaimonia.* Note that in this evolutionary tale, the strategies themselves form a population, which evolves as a *dynamical system.* If a given strategy spreads so as to engulf the entire population, it is an *evolutionarily stable strategy* (Maynard Smith and Price 1973) if its monopoly cannot be overthrown by a small scale invasion of newcomers who follow a different "mutant" strategy.[10] Should the dynamical system converge to an evolutionarily stable strategy, then the system is at equilibrium and all in the agent population follow a corresponding Nash equilibrium of the game.[11] The actual dynamics which governs the evolution of strategies depends of course upon the agent population. Biologists frequently use the *replicator dynamics* to model the spread of behaviors within nonhuman species. With this dynamics, the share of the population using a particular strategy grows in direct proportion to the current expected payoff of this strategy.[12] To be sure, the replicator dynamics is simpler than the actual dynamics of this evolution. The dynamics

[10] Game theorists usually speak of an evolutionarily stable strategy rather than an evolutionarily stable population. More precisely, a strategy s^* is evolutionarily stable if there is a constant ϵ, $0 < \epsilon < 1$, such that if the share of the population following s^* is $1 - \epsilon$ or greater and the rest follow any strategy $s' \neq s^*$, then s^*has a higher expected payoff against the whole population than s' has against the whole population. In other words, if an invasion of agents who follow a "mutant" strategy have an initial share in the population of ϵ or less, then the mutant strategy does worse on average than the incumbent strategy s^*, so that the mutant invaders will tend towards extinction. Equivalently, s^* is evolutionarily stable if:

(i) No strategy s has a higher expected payoff against s^* than s^* itself, and

(ii) If a strategy s' and s^*have an equal expected payoff against s^*, then s^*has a higher expected payoff against s' than s' has against itself.

These conditions capture the intuition that an evolutionarily stable strategy always does better against any mutant strategy than the mutant strategy does against itself.

[11] The converse does not hold, that is, not every Nash equilibrium corresponds to an evolutionarily stable strategy.

[12] If $p^t(s)$ is the proportion of strategy s in the population at time t, $u^t(s)$ is the average payoff of s at t and $\boldsymbol{u}^t$ is the average payoff for the whole population at t, then the replicator dynamics is defined by the differential equation

$$\tfrac{dp^t(s)}{dt} = \tfrac{p^t(s)}{\boldsymbol{u}}(u^t(s) - \boldsymbol{u}^t)$$

where the initial conditions are the proportions of strategies at time $t = 0$. A close approximation of this dynamics is given by the difference equation

$$p^{t+1}(s) = p^t(s)\tfrac{u^t(s)}{\boldsymbol{u}^t}$$

where the population is modeled as evolving at successive generations of discrete time.

which govern cultural evolution in human societies are surely even more complex. Still, some authors, including Skyrms, use the replicator dynamics as an admittedly crude approximating model of cultural evolution because it captures qualitatively the property which characterizes evolutionary systems, namely, the share of a type of individual in the system tends to increase monotonically with this type's fitness, by any measure of fitness. In any event, as Skyrms points out (p. 11), for many of his arguments the details of the dynamics are unimportant. What matters is which states are *attracting points* of evolution, since these points characterize the conventions that can emerge over time in a society.

Skyrms is not the first to apply evolutionary game theory to problems of distributive justice like the bargaining and ultimatum games, but he is the first to explore some of the key subtleties of the evolutionary approach. Suppose the members of a population have all settled into a strategy of claiming half the resource in the Figure 2 game. Then if some newcomers arrive who claim more than half, they will get nothing. If other newcomers arrive who claim less than half, they will get less than half. So a limited mutant invasion of a strategy other than claiming half does strictly worse than claiming half does against itself, and as a result the mutants will tend towards extinction. On the other hand, if all in a population claim any share other than half, an incoming mutant group who claim half will do better on average and drive the incumbent strategy to extinction. Hence claiming half is the unique evolutionarily stable strategy of the Figure 2 game. This would seem to point to a general explanation of the fair division norm: Evolution eventually carries any population to this norm.

However, Skyrms detects a serious flaw in this argument. An evolutionarily stable strategy is an attracting point of the dynamics of the evolutionary process, but not all such attracting points are evolutionary stable strategies. Suppose two thirds of the population claim $\frac{11}{12}$ and the remaining third claim $\frac{1}{12}$. Given the random pairing assumption, a $\frac{11}{12}$-claimer meets $\frac{1}{12}$-claimers and gets $\frac{11}{12}$ one third of the time, and gets nothing the rest of the time. A $\frac{1}{12}$-claimer gets $\frac{1}{12}$ in every encounter. So the average payoff for everyone in the population is $\frac{5}{4}$, and on average $\frac{11}{12}$ of the resource goes to waste. But this is also a stable point of evolution! The ratio of $\frac{11}{12}$-claimers to $\frac{1}{12}$-claimers leaves all with the same average payoff, that is, this *polymorphic* system is at equilibrium.[13] If any mutants enter into this population claiming less than $\frac{1}{12}$, they always get less than $\frac{1}{12}$. If mutants arrive claiming more than $\frac{11}{12}$, they always get nothing. And if mutants arrive claiming more than $\frac{1}{12}$ but less than $\frac{11}{12}$, then they get nothing two thirds of the time and their claim only one third of the time. No matter what strategy a mutant group might introduce into this polymorphic population, it will do worse on average than the incumbent population and will therefore be driven to extinction. So the evolutionary process could get caught in this *polymorphic trap* (Sugden 1986 pp. 68-69, Skyrms pp. 11-16).[14] Moreover, there are infinitely many other such polymorphic traps for this system, all of which leave all in the population worse off on average than they would be at the equal division norm. In computer simulations of populations which evolve via the replicator dynamics and in which the pure strategies are to demand some of

[13] In a population of $\frac{1}{12}$-claimers and $\frac{11}{12}$-claimers, $\frac{1}{12}$-claimers will get $1 + \frac{1}{12} \cdot 3 = \frac{5}{4}$ in every encounter while $\frac{11}{12}$-claimers will get a positive payoff exactly when they encounter $\frac{1}{12}$-claimers. If the fraction of $\frac{11}{12}$-claimers decreases below two thirds of the population, then the average payoff to a $\frac{11}{12}$-claimer becomes greater than $\frac{1}{12}$ because a $\frac{11}{12}$-claimer is meeting $\frac{1}{12}$-claimers more than one third of the time. Hence the $\frac{11}{12}$-claiming strategy is more reproductively fit than the $\frac{1}{12}$-claiming strategy, so the proportion of $\frac{11}{12}$-claimers would tend to increase. On the other hand, if the fraction of $\frac{11}{12}$-claimers increases above two thirds of the population, then the average payoff to a $\frac{11}{12}$-claimer falls below $\frac{1}{12}$ because a $\frac{11}{12}$-claimer is meeting $\frac{1}{12}$-claimers less than two thirds of the time. Now claiming $\frac{11}{12}$ is less reproductively fit than claiming $\frac{1}{12}$, so the proportion of $\frac{11}{12}$-claimers would tend to decrease. So the evolutionary dynamics would always drive a shift in the relative proportions of strategies back to two thirds $\frac{11}{12}$-claimers, one third $\frac{1}{12}$-claimers.

[14] In his analysis of a similar fair division game in sections 4.4 and 4.6 of *The Economics of Rights, Cooperation and Welfare*, Sugden (pp. 68-69) derives the set of all the evolutionarily stable equilibrium points in mixed strategies and observes that with the exception of the fair division equilibrium, each corresponds to a polymorphism in which some proportion of the population claim c of the good and the rest claim $1 - c$, $0 \leq c < .5$. Skyrms independently rediscovered the polymorphic trap problem in fair division games and was the first to show that evolutionary dynamics carries a population to a suboptimal polymorphism for a positive proportion of the possible initial conditions. Sugden does not try to explain the prevalence of the fair-division norm in light of the existence of polymorphic traps. Instead, Sugden argues that a variety of division norms can emerge if the symmetry of the fair division game is broken in various ways. As discussed below in §2, Skyrms considers some similar implications of symmetry breaking in Chapter 4 of *Evolution of the Social Contract.*

ten equal shares of a resource, Skyrms found that the populations converged to the fair division norm 62% of the time, but the rest of the time settled into one of the polymorphic traps (pp. 15-16).[15]

Skyrms' computer simulations suggest that while the fair division norm has a large *basin of attraction* of the evolutionary process, polymorphisms of unfair division pose a serious obstacle for the evolutionary argument for justice. Can one strengthen this argument so as to ensure that evolution always produces justice? One might introduce the possibility of large scale shifts in the population into the analysis. By definition, a stable point of the evolutionary dynamics can repel a small scale invasion or "rebellion" of mutants. But random fluctuations in the population might occasionally lead to a large scale rebellion. Even a stable equilibrium need not survive the forces of evolution forever. If large scale fluctuations are possible, albeit rare, then any equilibrium can ultimately be upset, and when the system converges again the result might be a different equilibrium. The potential for equilibrium transition is an important consideration in evolutionary game theory, but it does not by itself strengthen the case for the evolution-implies-justice argument.[16] If evolution can dethrone a stable polymorphism, it can also dethrone the fair division norm.

However, Skyrms argues that the evolution-implies-justice argument can be strengthened by allowing for another possibility, namely, that strategy encounters might be *correlated.* Evolutionary game theory traditionally assumes that strategy encounters are completely random, partly because this assumption is mathematically convenient. However, there are good reasons for supposing that like individuals in an evolutionary system have some tendency to encounter each other more often than random chance

[15] Skyrms ran 10,000 simulations of populations in which the initial proportions of strategies were chosen at random.

[16] This phenomenon *does* strengthen the argument if one assumes that the system evolves according to a dynamics for which the equal division norm has an especially large basin of attraction. See Foster and Young (1990) and Young (1993).

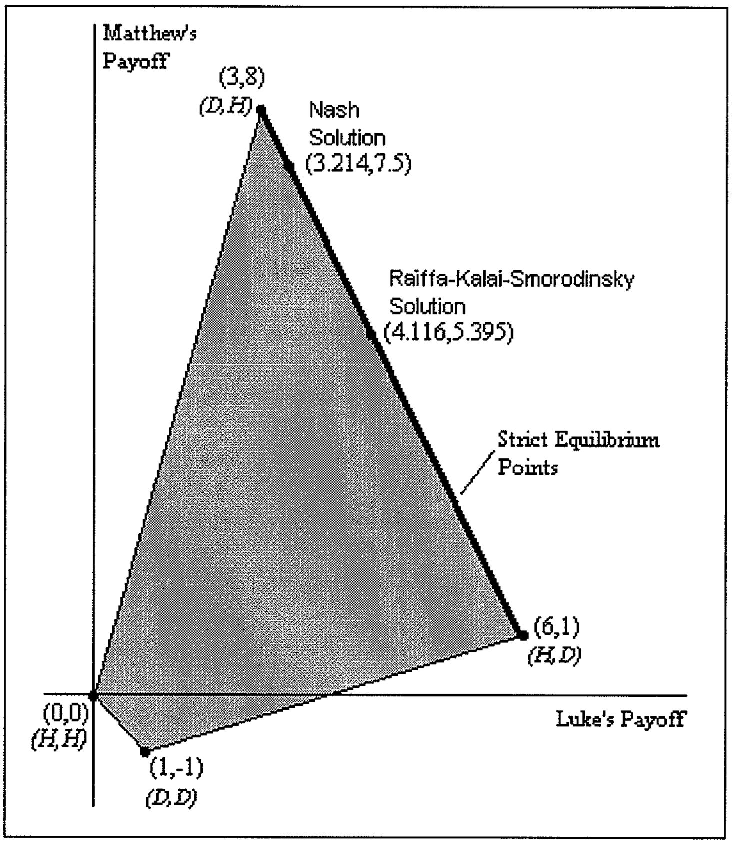
would predict. For instance, a group of mutant invaders or rebels might infiltrate the system as a "cluster" in close proximity to each other. Such mutants would meet each other more often than they would if they were randomly spread throughout the population. In the case of perfect correlation, like strategies always encounter like strategies, and in this case the only strategy in the Figure 2 game that can persist is to claim half. To assume perfect correlation is just as unrealistic as to assume no correlation at all. Still, Skyrms gives evidence from computer simulations that only a small amount of positive correlation in strategies in a bargaining problem will suffice to destabilize the polymorphic traps, so that the system stays at the fair division norm almost all of the time.

As for the ultimatum game, both Skyrms and Binmore point out that unlike bargaining problems, situations which have the structure of the one-shot ultimatum game seldom occur in real life. Still, Skyrms draws some striking conclusions from an evolutionary analysis of the ultimatum game. In the evolutionary model, agents sometimes enter into the system as proposers and other times as receivers, so each agent's strategy must determine a course of action in either role. Again, Skyrms uses computer simulations of evolution by the replicator dynamics of a simplified version of the base game. Rational choice game theory predicts that a rational agent will not adopt the fairness-norm strategy of offering half and accepting half, rejecting less than half. Similarly, a fairness-avoiding strategy which makes and accepts unfair offers but rejects fair offers violates the logic of rational decision. Nevertheless, Skyrms found that when the initial proportions of strategies in the population are chosen at random, the system always converges to a polymorphism containing either some fairness-norm agents or some fairness-avoiding agents. So evolution need not drive the fairness-norm strategy in the ultimatum game to extinction. On the other hand, evolution need not preserve the fairness-norm in this game, either. Perhaps most surprisingly, from almost any set of initial conditions, the replicator dynamics permits strategies that are mistakes from the perspective of rational choice game theory to persist in the population.[17]

Now let us return to the bargaining game. The Aristotelian norm of distributive justice allots equal shares to equals. The evolutionary argument for the equal division norm in the Figure 2 game tacitly assumes that the agents have no differences relevant to the good to be distributed. What if the agents are different in some relevant way? Aristotle maintains that distributive justice must be proportionate according to the relevant criteria. But whatever the relevant criteria, one needs an account of proportionality to determine what constitutes a just distribution. Braithwaite used the bargaining problem to analyze proportionality. Relevant differences between individuals can be reflected by assigning them *asymmetric payoffs*. Suppose that Luke's and Matthew's payoffs in another resource division problem are given in Figure 3.

[17] In a related study which Binmore does not discuss in detail in *Just Playing*, Binmore, Gale and Samuelson (1995) investigate the evolution of behavior in ultimatum games in which the agents are divided into two subpopulations, one of allocators and the other of recipients. They conclude that behavior which violates rational choice game theory can evolve given sufficient "noise" in the evolutionary system. (Mutation is one form of "noise" in evolution.)

Figure 3. Fair Division Problem With Asymmetric Payoffs



The asymmetries in this game appear to favor Matthew. The challenge is determining what constitutes proportionate shares in light of these asymmetries. Nash (1953) argued that any bargaining problem can be analyzed either from the perspective of an arbiter, who employs certain axioms to determine a just division, or from the perspective of the agents themselves, who try to resolve the problem via some bargaining process. Braithwaite, and Gauthier (1986) after him, took the axiomatic approach. They defend a solution proposed by Raiffa (1953) and Kalai and Smorodinsky (1975) which divides the resource so that if the payoff for each is 0 at the (H, H) outcome, then the ratio of the payoff an agent receives to his payoff if he gets all of the resource is the same for each agent.[18] But as Luce and Raiffa admit (1957, pp. 145-150), the arguments for the Raiffa-

[18] Raiffa introduced this solution to the two-agent bargaining problem by example, and Kalai and Smorodinsky gave the axioms for this solution. Raiffa and Kalai and Smorodinsky present this solution to the bargaining problem without arguing that it is necessarily the "correct" solution. Braithwaite was the

Kalai-Smorodinsky solution are by no means decisive. Nash proposed a different axiomatic solution for the two-agent bargaining problem, according to which the resource is divided so as to maximize the product of the two agents' payoffs given that they both receive 0 if both follow H.[19] A utilitarian solution would divide the resource so as to maximize the sum of the payoffs.[20] And there are still other axiomatic solutions defended in the literature. There is no generally accepted solution to the general two-agent bargaining problem, let alone the general bargaining problem for two or more agents.

Skyrms applies evolutionary arguments to the asymmetric bargaining problem. All of the axiomatic solutions to the two-agent bargaining problem are strict Nash equilibria. One might hope that evolution will always select the same equilibrium in bargaining games. Skyrms shows by example that this is not the case for populations that evolve according to the replicator dynamics. In different bargaining games with asymmetric payoffs and in which the agents' strategies are to claim some discrete share of the good at stake, the Nash bargaining solution emerges most frequently in computer simulations, but a variety of other equilibria also emerge, including the Raiffa-Kalai-Smorodinsky solution and other equilibria not prescribed by any axiomatic solution. Skyrms tentatively concludes that evolution might produce a far greater variety of norms for distributing goods than any single axiomatic bargaining theory would predict (p. 107).

first to argue that the Raiffa-Kalai-Smorodinsky solution defines the just distribution in a resource division problem. In the two-agent case, the Raiffa-Kalai-Smorodinsky solution is Gauthier's *minimax relative concession* solution, which Gauthier (1986, Chapter V) formulates for the general n-agent bargaining problem.

[19] As stated, these definitions of the Nash and the Raiffa-Kalai-Smorodinsky solution presuppose that the agents' payoffs are scaled that both agents get a payoff of 0 if both carry out threats to follow H and therefore get none of the resource. This involves no loss of generality, since payoffs are unique only up to a choice of scale. (See note 6.) One can also give slightly more complex definitions of the Nash and the Raiffa-Kalai-Smorodinsky solutions that do not assume that the *threat point* of the bargaining problem is at the origin.

One of the axioms of both the Nash and the Raiffa-Kalai-Smorodinsky solutions is a symmetry axiom which requires that agents receive equal shares if their payoffs are completely symmetric. This symmetry axiom is of course equivalent to the equal division norm.

[20] In the Figure 3 game, utilitarianism recommends that Matthew gets all of the resource!

This agrees with the claims of some communitarian philosophers who, following Walzer (1983), deny that there is a single rule for distributive justice that applies to all circumstances. But the Nash solution appears to evolve more frequently than the other axiomatic bargaining solutions. Skyrms suggests that philosophers who have focused on the relative merits of utilitarianism and the Raiffa-Kalai-Smorodinsky solution might take the Nash solution more seriously.

As Skyrms acknowledges, the analysis of fair division problems in *Evolution of the Social Contract* leaves a great deal of territory under-explored. He limits his discussion to two-agent fair division problems because there is no developed account in game theory of how *coalitions* form in games of three or more agents. Much work remains to be done both in developing computer models less simplistic than the models Skyrms uses in his simulations, and in showing analytically under what conditions a fair division norm will evolve and persist. Additionally, Skyrms does little to connect his general evolutionary argument with the history of norms of distributive justice in human societies. Skyrms' main achievement in this part of *Evolution of the Social Contract* is a conceptual breakthrough which should spur a body of continuing research. By allowing for the possibility of polymorphic pitfalls in the bargaining problem, Skyrms gives a powerful rebuttal to those who naively conflate the idea behind the evolutionary slogan "survival of the fittest" with social optimality. By relaxing the unrealistic assumption of completely random matching, he introduces a much more complex evolutionary game theory, which can explain how justice evolves better than the traditional theory.[21]

[21] On p. 58 of *Evolution of the Social Contract*, Skyrms considers the possibility of *negative* or *anti-correlation* in evolutionary game theory. An *"intersection game"* is a 2-agent game in which each agent receives a positive payoff if one follows the pure strategy S ("stop") and the other follows the pure strategy G ("go"), and otherwise they both receive 0. Skyrms shows that if the members of a population who engage in an intersection game are more likely to encounter others who follow strategies *different* from their own than random chance would predict, then they can achieve a greater average payoff than they would achieve in the absence of correlation. One line of research some of Skyrms' readers are already exploring is how anti-correlation can affect the evolution of behavior in resource division games. D'Arms (1996) and D'Arms, Batterman and Gorny (1997) argue that if certain strategies in a bargaining game are for some reason anti-correlated, this would tend to increase the stability of some of the polymorphic traps.

§2. Correlated Convention

> *What is a man's property?* Anything which it is lawful for him, and for him alone, to use. *But what rule have we, by which we can distinguish these objects?* Here we must have recourse to statutes, customs, precedents, analogies, and a hundred other circumstances; some of which are constant and inflexible, some variable and arbitrary.
>
> David Hume, *Enquiry Concerning the Principles of Morals*

In the symmetric bargaining game, the unique evolutionarily stable strategy is to follow the equal division norm, which yields the one symmetric division of the good. A variety of other problems seem to call for an asymmetric solution. For instance, in the "Chicken" game of Figure 4, either of the two strict Nash equilibria (D, H) and (H, D) leaves one of the agents strictly worse off than the other.

Figure 4. Chicken[22]

		Agent 2	
		D	H
Agent 1	D	$(3,3)$	$(2,5)$
	H	$(5,2)$	$(0,0)$

$D =$ dove, $H =$ hawk

These works obviously imply that when one applies evolutionary game theory to any particular biological or cultural context, one must consider what kinds of correlation make sense in this context.

[22] This game is known as *Chicken* because of a motivating story inspired by the ending of the film *Rebel Without A Cause*: Two individuals drive towards each other in their automobiles. Each driver can either stay the course (H) or "chicken out" and swerve away from the other at the last moment (D).

This game models situations which are resolved when exactly one agent “takes charge”, as when agents are trying to decide who owns a good that cannot easily be divided or who will have to perform some important task. Chicken resembles the Figure 1 game in that each would most like to follow H while the other follows D. However, in this game if both follow H a conflict erupts, which is the outcome each most wants to avoid. Exactly one of them must give in and choose D, but the symmetry of the problem seems to leave us no clue as to which agent will give in.

The fundamental problem of game theory is the equilibrium selection problem. In most cases, there is no *a priori* reason to suppose that rational agents will follow an equilibrium at all, let alone any particular equilibrium.[23] If Agents 1 and 2 are both rational and are engaged in a game characterized by the payoffs of Figure 4 and even have common knowledge[24] of all this, it does not follow that the strategies they choose will form an equilibrium. If their common knowledge is augmented so that the *strategies* each intends to follow are also common knowledge, this is a sufficient condition for these strategies to constitute an equilibrium.[25] Binmore aptly defines a community's *culture* as its store of common knowledge, and argues that the members of any community apply the common understandings of their community to pick out equilibria in the various games in which they engage each other (pp. 270-271). But how do individuals come to have this common knowledge? How do they come to follow one equilibrium rather than any other equilibrium?

[23] An exception is a game that is *dominance solvable*, that is, a game with a unique Nash equilibrium which each agent can predict the others will follow by stepwise ruling out strategies that no rational agent will follow. The Prisoners’ Dilemma game given in Figure 5 is dominance solvable.

[24] David Lewis gave the first precise analysis of common knowledge in *Convention*. According to Lewis, a proposition A is common knowledge for a group of agents if each agent knows that all know A and knows that anything she can infer from this, all can infer. Lewis-common knowledge implies an alternate account of common knowledge Schiffer (1972) gave independently: A is common knowledge for a group of agents if each agent knows A, each agent knows that each agent knows A, and so on, *ad infinitum*. Schiffer’s definition has become the standard definition of common knowledge in the literature.

[25] See Aumann (1987) and Brandenburger and Dekel (1988) for proofs of this result.

One might naturally suppose that the agents can select an equilibrium to follow if they can discuss their situation before they act. Binmore follows a number of authors who stress the importance of such "cheap talk" in equilibrium selection problems. However, the cheap talk explanation has its limitations. Binmore argues that cheap talk will be effective in helping agents select an optimal equilibrium only if there is no ambiguity as to which is the optimal equilibrium to aim for (p. 209).[26] Even in games as simple as Chicken, there is such ambiguity. Moreover, to appeal to communication as an equilibrium selection mechanism is to introduce some preexisting higher-order coordination into the analysis. As Lewis (1969, Chapter IV) argues, the conventions of language themselves are equilibria of games. If within a group it is common knowledge that when one sends a given signal the receivers react appropriately with respect to a state of affairs, then for this group this signal *means* this state obtains. Such a *meaning convention* is a Nash equilibrium of a corresponding *signaling game*. Meaning conventions present us with another equilibrium selection problem. Why does one meaning convention emerge within a group when any of a number of different meaning conventions might have emerged?

Quite early in the history of game theory, Luce and Raiffa suggested another explanation of equilibrium selection in their *Games and Decisions*: Agents gradually settle into an equilibrium as the result of some trial and error process (1957, p. 105). Luce and Raiffa echoed Nash's argument that his equilibrium concept has a "mass-action" interpretation, according to which each agent in a game chooses a strategy which is a best response to what she expects the others to do given the accumulated data from past plays, so that in the limit the agents follow an equilibrium (1996, pp. 32-33).[27] David Hume

[26] An outcome of a game is *(Pareto) optimal* if no other outcome yields any agent a higher expected payoff and none a lower expected payoff.

[27] Nash's 'Noncooperative games' (1951) is a revised version of his doctoral thesis, submitted to the Princeton University department of mathematics in 1950, with one section missing entitled *Motivation and interpretation.* This section of Nash's thesis contains his discussion of the "mass-action" interpretation

articulated the basics of this idea much earlier in *A Treatise of Human Nature*, where he argued that a convention gradually emerges by "a slow progression" of trial and error learning (1740, p. 490). However, most game theorists did not actively explore this kind of dynamical explanation of equilibrium until the 1980's, in large part because we have a very limited understanding of the dynamics of inductive learning. The appearance of Maynard Smith and Price's evolutionary game theory helped stimulate a now burgeoning body of research on dynamical models of learning in games, because one can view the spread within a population of strategies agents have learned to follow as a kind of evolution. In earlier work, Binmore (1987*a*, 1987*b*, Binmore and Samuelson 1992, Binmore and Samuelson 1996) and Skyrms (1990,1991) have made significant contributions to this literature. In their new works, they consider what dynamical models of equilibrium selection can tell us about the emergence of social norms.

One can define a *convention* or *norm* for a given community either as a system of reciprocal expectations that all will follow one of several distinct equilibria (Lewis 1969, Sugden 1986, Bicchieri 1993), or as a rule requiring one to follow this equilibrium that all expect to be followed (Binmore, p. 271). In the Chicken game, two obvious candidates for norms are the rules "Follow (H, D)." and "Follow (D, H)." However, it would appear that neither of the strict Nash equilibria of Chicken is a stable point of evolution. For if D-followers predominate the population, then H is the better strategy, and vice versa. The only evolutionarily stable state of a population of H and D strategies is a polymorphism in which half of the population follow D and the other half follow H.[28] The symmetry of the Chicken game seems to inhibit the evolution of a strict Nash equilibrium. In addition, the results of evolution do not seem to cohere well with the definition of norms in terms of equilibrium.

of the Nash equilibrium concept. A recent (1996) volume of reprints of Nash's game-theoretic essays reproduces this section.

[28] This polymorphism corresponds to the mixed Nash equilibrium of the Figure 4 game.

Like Sugden before him, Skyrms argues that the dynamics of evolution actually *breaks* symmetries, so that in the end evolution can select any strict equilibrium. Random recombination and mutation continually perturbs a biological system. Occasionally, sufficiently many perturbations arise within the system to upset "nature's balance". A human culture is similarly bombarded by behaviors which deviate from established behavioral patterns. Whether one regards those who deviate as innovators or as rebels, they might sometimes lead society towards a new norm. If the members of a community are stuck in the polymorphism in which half of them are following D and the remaining half are following H in the Chicken game, then agents who follow a slightly more complex strategy can disrupt the symmetry of the situation. Suppose that into the polymorphism a group enters whose members peg their strategies on what seems to be an extraneous bit of information. When any member of this group is about to engage someone in a Chicken encounter, he waves at a bystander they both can see, and then follows H only if the bystander waves directly back at him (and not the other). A bystander either returns Agent 1's wave or Agent 2's wave, or does not clearly respond exclusively to either agent, and each of these possibilities occurs equally often in the community. So from the perspective of an agent in this select group, the bystander's response is a random event with three equally likely outcomes. But surprisingly, the external event guides the strategies of the group members so as to undermine the polymorphism. Group members achieve the same expected payoff against those who follow D or H unconditionally as unconditional strategists do against themselves, and they achieve a strictly greater expected payoff when paired against each other.[29] This *conditional* strategy is evolutionarily stable, and will tend to overwhelm the

[29] The expected payoff of following either H or D unconditionally in the polymorphism is $2\frac{1}{2}$, since an unconditional H- or D-follower meets H-followers half the time and D-followers the other half of the time. Against unconditional strategists, a group member follows D two thirds of the time and H the remaining third of the time, to achieve an expected payoff of $\frac{2}{3}(\frac{1}{2}\cdot 3+\frac{1}{2}\cdot 2)+\frac{1}{3}(\frac{1}{2}\cdot 5+\frac{1}{2}\cdot 0)=2\frac{1}{2}$. Against each other, group members alternate between (D,H), (H,D) and (D,D) one third of the time each, to achieve an expected payoff of $\frac{1}{3}\cdot 2+\frac{1}{3}\cdot 3+\frac{1}{3}\cdot 5=3\frac{1}{3}$.

polymorphism. If this happens, then the community members are following a *correlated equilibrium* of the Chicken game. Robert Aumann formulated a correlated equilibrium concept for game theory in a pair of seminal articles (1974, 1987). Correlated equilibrium generalizes the Nash equilibrium concept by relaxing Nash's requirement that the agents follow probabilistically independent strategies. The rule "Follow H if the bystander waves at you alone, and D otherwise." determines a strict equilibrium in which the agents' strategies are correlated with what they observe the bystander doing. Skyrms is perhaps the first author to clearly draw the connection between evolutionary game theory and correlated equilibrium.

To be sure, in this evolutionary story one correlated equilibrium emerges as the equilibrium which determines the community norm when many other correlated equilibria might have emerged. Individuals in the Chicken game could follow a correlated equilibrium in which one follows H only if the bystander waves to the *other* agent only. If Chicken encounters take place when agents randomly meet going in different directions at an intersection, they can follow an even better correlated equilibrium if they follow the rule "Follow H only if the other is to my left."[30] Any of a number of environmental factors can cue the agents' actions so they follow an equilibrium. As David Hume observed, in many cases a "finders-keepers" convention settles the problem of property acquisition (1740, pp. 503-505). In some other contexts, property is acquired according to rules of primogeniture. Skyrms regards this general phenomenon as a chief advantage, perhaps *the* chief advantage, of using evolutionary game theory as a tool in social philosophy.

> When we investigate this interactive [evolutionary] dynamics we find something quite different from the crude nineteenth-century determinism of the social Darwinists on the one hand, and Hegel and Marx

[30] At this correlated equilibrium, the agents alternate between (D, H) and (H, D) with probability one half each, achieving an expected payoff of $\frac{1}{2} \cdot 2 + \frac{1}{2} \cdot 5 = 3\frac{1}{2}$.

> on the other. It is apparent, even in the simple examples of this book, that the typical case is one in which there is not a unique preordained result, but rather a profusion of possible equilibrium outcomes. The theory predicts what anthropologists have always known --- that many alternative styles of social life are possible. (*Evolution of the Social Contract*, p. 109)

Binmore gives concurring discussion in *Just Playing*. Thomas Huxley (1888) and Henry Sidgwick (1902) were among the first to point out that a popular version of ethical evolutionism relies on a variant of the naturist fallacy: Human communities inevitably evolve into a particular form of society, so we should strive to achieve this society.[31] Sidgwick (pp. 136-137) reminds us that a defender of such *final state evolutionism*[32] illegitimately draws a moral conclusion from an allegedly scientific premise. Skyrms and Binmore reject the "scientific" premise of final statism in light of evolutionary game theory. In general, games have a diversity of evolutionarily stable equilibrium points which correspond to the different sets of social norms that can emerge in different societies. This point has not always been fully appreciated, perhaps because Axelrod's (1984) and Gauthier's (1986) landmark works gave some the impression that game theory prescribes a unique set of norms of social cooperation. Binmore's and Skyrms' new works should clear up this fundamental misunderstanding once and for all. From the game-theoretic perspective, a culture can evolve into one of many possible *social contracts*, which Binmore defines as the set of all the conventions which prevail in society (p. 5, p. 271). Other traditions in moral philosophy allow for the specifics of norms to vary at least to some extent across cultures. Evolutionary game theory provides a framework for *explaining* such variation. Moreover, since the payoffs in an evolutionary game reflect the good of the individuals who engage in the game,

[31] The general inference pattern of the naturist fallacy is: We have a natural tendency towards being X, so we ought to encourage being X.

[32] My term, not Sidgwick's.

evolutionary game theory can explain the variation in norms across cultures as alternative ways in which cultures encourage and achieve the same good. Finally, since the magnitude of the payoffs reflects the relative desirability of alternative outcomes for agents, evolutionary game theory even provides a framework by which one can argue that some social states are more desirable than others. Hence evolutionary game theory indicates not only that cultural evolution is *not* deterministic, but that it can make sense to argue that a society should abandon a current norm in favor of a better norm.

Skyrms (Chapter 5) applies this framework to an important special case: the conventions of language. Russell (1921), and Quine after him (1936), argued that there are no linguistic conventions, except perhaps some set up by formal agreement, because they thought that agents would need language in order to establish a convention of language. Lewis (1969) countered that a convention of meaning is established when agents have common knowledge that they follow a signaling system of an appropriate signaling game, and since agents do not necessarily require "cheap talk" to pick an equilibrium, the circularity that worried Russell and Quine is avoided. Lewis concurred with Schelling (1960) that agents can sometimes select an equilibrium without prior communication if this equilibrium is somehow *focal* or *salient*. However, Skyrms argues that like the cheap-talk explanation, the salience explanation of equilibrium selection has shortcomings. Skyrms shows by example that in some signaling games, no signaling system is clearly salient. Indeed, in general a signaling game has a number of Nash equilibria that are not even signaling systems. Skyrms provides an alternate explanation of the origins of meaning conventions. He shows that the set of signaling systems in any signaling game coincides exactly with the evolutionarily stable strategies of this game. Meaning conventions therefore are the result of the dynamics of evolution.

Of course, this analysis of coordinated behavior in society via evolutionary game theory raises a crucial question: How did the members of a certain "mutant" group learn to follow the correlated equilibrium that ultimately spread throughout the population? A wealth of experimental studies confirm that people settle into an equilibrium of a coordination game far more readily than chance would predict.[33] But the specifics of the learning processes by which individuals reach equilibrium are not well understood to date. While game theorists have made significant progress in analyzing learning in games, many elementary problems remain unsolved.[34] To give only one example, learning models that can explain how agents constantly follow one Nash equilibrium of a coordination game like Chicken cannot account for cases in which agents take turns between (H, D) and (D, H), and yet people frequently learn to follow a "taking turns" equilibrium in Chicken-type games both in the laboratory and in real life. While Skyrms (1990) has presented some formal models of learning in games of his own, he is aware that we presently have but a crude understanding of this learning. In *Evolution of the Social Contract*, Skyrms limits himself to qualitative observations on how agents can learn to follow a correlated equilibrium. Each agent will enter into a game with a set of beliefs regarding which combination of strategies the other agents will follow. Skyrms assumes that agents will modify these beliefs, quantified as probability distributions, recursively according to the frequency of strategy profiles chosen in the past. If the system of beliefs converges, the limit point will be an equilibrium which the agents have learned inductively to follow. Most of the learning models in the literature have assumed, for mathematical convenience, that the agents' belief distributions are such that they expect each other to choose probabilistically independent strategies. This assumption guarantees that the limit of inductive learning will be a Nash equilibrium. Skyrms argues that if one relaxes the probabilistic independence assumption, then inductive learning can converge to correlated equilibrium. Indeed, the process of inductive learning itself can

[33] Schelling recounts the striking results of his early experimental studies of coordination problems in Chapter 3 of *The Strategy of Conflict*. Kagel and Roth (1995) and Crawford (1997) survey most of the relevant experimental results known to date.

[34] Fudenberg and Levine (1998) have recently published an important work which surveys and compares most of the learning models in the existing literature.

generate correlation in beliefs, so that even if their beliefs initially satisfy probabilistic independence the agents can still eventually converge to a correlated equilibrium. This is an important step forward, but only a first step. Game theorists in the future will have to address many questions raised by Skyrms' suggestion, including: (1) What is a correct model, or set of models, of correlated inductive learning?, and (2) Under what general conditions will such inductive learners converge to an equilibrium of the game?

§3. Reciprocal Aid

When each member of a community contributes towards a common good, all are better off for it. But if the common good benefits each individual only to the degree that the others contribute, then to contribute is to bear a personal cost to no personal advantage. So individuals will tend not to contribute towards the common good, even when they realize there will be no common good if no one contributes. This *free-rider* argument has a distinguished history in moral and political philosophy.[35] The logic of this argument for a community of two is captured by the game presented in Figure 5.[36]

[35] Here is a partial citation list of relevant passages in philosophical classics: Aristotle, *Politics* 1261b32, Aquinas, *Summa Theologiæ* II-II, Question 66, Second Article, Hobbes, *Leviathan*, Chapter 17, Hume, *A Treatise of Human Nature*, Book III, Section VII, Rousseau, *The Social Contract*, Book II, Chapter III and Mill, *Principles of Political Economy*, Book V, Chapter XI, Section 12.

[36] At one time, it was thought that all free-rider problems have the structure of a multi-agent Prisoners' Dilemma. Taylor and Ward (1982) and Hampton (1986, 1987) were among the first to argue that for three or more agents, free-riding can occur in a wider class of games than the Prisoners' Dilemma. One might regard the n-agent Prisoners' Dilemma, in which D is each agent's strictly dominant strategy, as an endpoint case, in the sense that if one can explain why agents would ever cooperate in this Prisoners' Dilemma, one can explain cooperation in all of the n-agent free-rider problems.

Figure 5. Prisoners' Dilemma

		Agent 2	
		C	D
Agent 1	C	$(2, 2)$	$(0, 3)$
	D	$(3, 0)$	$(1, 1)$

C = cooperate, D = defect

This is the Prisoners' Dilemma. Each agent can either *cooperate* (C) by contributing towards the common good, or *defect* (D[37]) by not contributing. To defect is each agent's unique best strategy in the Prisoners' Dilemma, so (D, D) is the only Nash equilibrium, even though both would be better off at (C, C). The Prisoners' Dilemma seems to throw the rationality of social cooperation into doubt. In particular, if to cooperate is to follow the norms of justice and thereby help sustain the peace in society, then justice evidently contradicts prudence.

David Gauthier recognized that a sound argument for the rationality of cooperating in the Prisoners' Dilemma would be a solution to the perennial problem of reconciling justice with self-interest. Gauthier's attempt to provide such an argument is the centerpiece of his *Morals By Agreement* (1986). Gauthier (Chapter VI) introduces and defends a *constrained maximization principle* as a rule of a rational choice in games. Constrained maximization recommends that an agent follow her end of a fair and optimal outcome in a game if she expects the other agents to follow this outcome and she benefits

[37] Not to be confused with the "dove" strategy in the Figure 1 and Figure 4 games.

as a result.[38] In the 2-agent Prisoners' Dilemma, a constrained maximizer cooperates just in case she believes she is paired with another constrained maximizer. The core of Gauthier's argument for constrained maximization is deceptively simple: In problems with a Prisoners' Dilemma structure, a constrained maximizer does as least as well an agent who maximizes expected payoff "straightforwardly", and can do strictly better, so long as agents' dispositions to act as "straightforward" or constrained maximizers are sufficiently clear to each other. Unfortunately for Gauthier's program, constrained maximization suffers from insuperable difficulties. There is no clear way to apply the constrained maximization principle to cases, like Chicken, with a unique fair optimal outcome but multiple Nash equilibria.[39] A number of other authors, in particular Bicchieri (1993) and both Skyrms and Binmore (1994, 1998), argue forcefully that constrained maximization is an untenable principle of rational choice even in the games, like Prisoners' Dilemma, for which it was designed. To be a constrained maximizer is to adopt a disposition to conditionally cooperate when to defect is always one's unique best strategy whatever the others actually *do*. We seem to have no way of reliably detecting who adopts such a disposition. And what could make such a disposition credible? Apparently, to make her constrained maximizer disposition believable to others, an agent must in effect be able to bind herself to cooperate under certain circumstances. But if agents can somehow restrict their pure strategy options this way, then the game does not

[38] These are sufficient, but not necessary, conditions for a constrained maximizer to choose to cooperate in free rider problems. See pp. 167-177 of *Morals by Agreement.*

[39] It does not even follow from Gauthier's official definition of constrained maximization (p. 167) that constrained maximizers who recognize each other as such will necessarily follow (C, C) in the Prisoners' Dilemma. However, Gauthier clearly intends constrained maximization to be understood to imply that agents who identify each other as constrained maximizers will cooperate when paired in the Prisoners' Dilemma.

In the case of Chicken, on a literal reading of Gauthier's official definition of constrained maximization, neither agent would deviate from either strict Nash equilibrium to the (D, D) outcome. But according to another characterization of constrained maximization (p. 170), a constrained maximizer in Chicken ought to follow her part of (D, D). This would seem to imply that an H-follower can always exploit a constrained maximizer in Chicken, because H is the best response to D. Hence this situation differs from that of the Prisoners' Dilemma, where Gauthier can, and does, argue that if a constrained maximizer expects the other agent to deviate from (D, D), H is the constrained maximizer's best reply.

really have a Prisoners' Dilemma structure in the first place. In the end, Gauthier's attempt to reconcile justice and self-interest via a "reformed" rational choice game theory fails.

Still, Gauthier's motivating insight is an important one: The apparent tension between justice and prudence is a tension between cooperative behavior and equilibrium behavior. But as Binmore puts it, philosophers who argue that a rational agent cooperates in the one-shot Prisoners' Dilemma are really giving "a wrong analysis of the wrong game (1994, p. 174)". Binmore argues that *repeated games* are the relevant games for analyzing problems of mutual aid in society, since life is in fact a series of repeated interactions. In an indefinitely long sequence of Prisoners' Dilemmas repeated over time, strategies which reward others' past cooperation with cooperation can be rational, even though to cooperate is never rational in the one-shot Prisoners' Dilemma. Binmore and Skyrms (1998) both acknowledge that David Hume presented an informal version of this argument in *A Treatise of Human Nature*. Hume realized that certain problems of mutual aid resemble the Prisoners' Dilemma in structure (1740, p. 520).[40] Hume also realized that to cooperate in one instance of a mutual aid problem could help establish a pattern of *reciprocal aid* over time.

> I learn to do a service to another, without bearing him any real kindness; because I forsee, that he will return my service, in expectation of another of the same kind, and in order to maintain the same correspondence of good offices with me or with others. And accordingly, after I have serv'd him, and he is in possession of the advantage arising from my action, he in

[40] Obviously, Hume did not know the technical vocabulary of game theory, but his analysis of convention in Part III of *A Treatise of Human Nature* contains a number of profound informal game-theoretic insights. Skyrms (1998) discusses some of Hume's game-theoretic arguments in detail. In the introduction to his game theory text *Fun and Games*, Binmore (1992, p. 21) calls Hume the "true founding father" of game theory.

> induc'd to perform his part, as forseeing the consequences of his refusal.
> (Hume 1740, p. 520)

The backbone of the theory of repeated games is a set of *folk theorems*, so-called because game theorists discussed these results informally for years before any of their proofs were published. The folk theorems tell us that any payoff vector in a one-shot base game that gives each agent more than the worst payoff the others can force her to accept can be achieved in an equilibrium of the corresponding indefinitely repeated game. When the base game is the Prisoners' Dilemma, the folk theorems tell us that there is an equilibrium in the indefinitely repeated game at which the agents are achieving the fair and optimal payoffs of mutual cooperation in the base game. Hume understood this in an informal way. Given an opportunity to contribute towards the common good, cooperating now can make sense, if this influences others to cooperate in return later.

An obvious corollary to the folk theorems is that an indefinitely repeated game typically has infinitely many equilibrium points. In the indefinitely repeated Prisoners' Dilemma, there are Nash equilibria of mutual cooperation, as Hume realized informally. However, there are many other Nash equilibria of this game. The agents are at another equilibrium of indefinitely repeated Prisoners' Dilemma if they defect every time. There are also Nash equilibria in which the agents follow C sometimes and D sometimes, including cases in which one agent is exploiting the other. Suppose Agent 1 cooperates 51% of the time so long as Agent 2 cooperates always, and if Agent 2 ever defects, Agent 1 then defects always. Then a best reply for Agent 2 is to cooperate always if Agent 1 sticks to his strategy, and defect always if Agent 1 deviates. To be sure, at this equilibrium Agent 1 enjoys an average payoff of $.51 \cdot 2 + .49 \cdot 3 = 2.49$ while Agent 2's average payoff is only $.51 \cdot 2 + .49 \cdot 0 = 1.02$, but Agent 2 puts up with the situation because to deviate would leave her worse off. In indefinitely repeated Prisoners

Dilemma, agents can follow an equilibrium of mutual cooperation, but from this fact we must not slide into concluding that they will follow such an equilibrium.

Indefinitely repeated Prisoners' Dilemma presents another equilibrium selection problem, one for which evolutionary game theory seems ideally suited. Axelrod pursued this idea in *The Evolution of Cooperation*. Axelrod argues that agents confronted with a sequence of indefinitely repeated Prisoners' Dilemmas will gradually settle into an equilibrium of mutual cooperation. The strategy which characterizes this equilibrium is *tit-for-tat*, that is, cooperate in the first round of the base game, and from then on do what the other did at the previous round. Axelrod gave three different arguments for this conclusion, one from Maynard Smith and Price's then new evolutionary game theory, one from historical evidence and a third from a striking set of computer simulations. Axelrod showed analytically that under fairly general conditions, a population of tit-for-tat-followers can repel any limited invasion of agents who ever exploit a C-move by replying with an D-move, and concluded that tit-for-tat is evolutionarily stable. Axelrod pointed to the surprising periods of informal truce along parts of the Western Front during the First World War as a case study of agents who sustained a mutually beneficial outcome with tit-for-tat behavior. Finally, Axelrod presented the results of an experiment, in which various individuals submitted strategies, in the form of computer programs, for playing the repeated Prisoners' Dilemma. These strategies were then matched against each other round-robin in a computer simulated sequence of Prisoners' Dilemma games. Of the sixty-three strategies submitted, tit-for-tat emerged the winner of Axelrod's tournament, in that tit-for-tat garnered the greatest total payoff of all the entries. Axelrod ran a subsequent series of computer simulations using the sixty-three tournament strategies, in which the proportion of each strategy in the population evolved according to a variation of the replicator dynamics. After 1000 generations, tit-for-tat again emerged as the most successful strategy. *The Evolution of Cooperation* has made a great popular impact. On the surface, Axelrod seems to have solved the old free-rider problem with evolutionary game theory.

Binmore goes to some effort to burst what he calls Axelrod's "tit-for-tat bubble (p. 317)". Binmore thinks that Axelrod's computer simulations may offer insights into how mutual aid might first have emerged in biological evolution (p. 320). One might say the same of Skyrms' analysis of mutual aid in *Evolution of the Social Contract* (Chapter 3), where he shows that C in one-shot Prisoners' Dilemma is evolutionarily stable in correlated evolutionary game theory, even though only D is evolutionarily stable in the traditional evolutionary game theory. However, Binmore argues that we should not conclude from Axelrod's arguments that human societies typically evolve into tit-for-tat societies. The number of strategies an agent can employ in a repeated Prisoners' Dilemma grows exponentially with the number of rounds. Not only does Axelrod limit his analysis to the strategies submitted to his tournament, but by running the tournament as a round-robin tournament, Axelrod glosses over the possibility that a tit-for-tat agent might encounter more than one kind of strategy over the sequence of repeated Prisoners' Dilemmas. Tit-for-tat does not really "win" Axelrod's subsequent evolutionary simulation, since it only forms the largest fraction of a polymorphism of the sixty-three original entries that survives after 1000 generations. Even this result occurs only because Axelrod assumes that all sixty-three strategies are equally represented in the first generation. So Axelrod's own simulations do not really confirm his thesis that tit-for-tat is evolutionarily robust (p. 314). As for Axelrod's anecdotal evidence from trench warfare, while it has seldom been noticed in the literature, the periods of informal truce are not really examples of tit-for-tat behavior. For one thing, as Binmore points out (p. 319), these spontaneous truces did not occur between sides that were initially cooperative.

For another, each side was evidently far more punitive when the other side violated a truce than the tit-for-tat strategy requires.[41]

Binmore also debunks the myth that Axelrod proved that tit-for-tat is evolutionarily stable (p. 320-323). For a mutant invasion of agents who cooperate unconditionally can successfully infiltrate a population of tit-for-tat agents, a fact that Axelrod failed to consider. And if the proportion of agents in the population who cooperate always grows sufficiently large, then another mutant invasion of agents who always defect can destroy the pattern of mutual aid. A tit-for-tat population can also be invaded by a mutant group who follow the *grim* strategy of cooperating until the other defects once, and then defecting always. More generally, for any pure strategy S of the repeated Prisoners' Dilemma, there is another strategy S' which treats both S and S'-followers as S-followers treat each other, and such that S and S' treat the followers of some other strategy differently. Since S- and S'-followers do equally well against each other as they do against themselves, a group of S'-following mutants can always invade a population of S-followers, and vice-versa. So no pure strategy for the indefinitely repeated Prisoners' Dilemma is evolutionarily stable.[42]

Binmore argues that we should not focus on any single strategy as being the "correct" way to explain cooperation in the repeated Prisoners' Dilemma (p. 322). Cultural evolution does not inevitably result in a pattern of mutual aid, and even when mutual aid prevails in a particular society, this is likely to be the result of several different

[41] While few seem to have noticed it, Axelrod observes this himself on p. 80 of *The Evolution of Cooperation.* Axelrod suggests that there was some "damping process" that mitigated the effects of retaliations, so that the retaliations did not provoke an escalation of hostilities. But even if there had been such a damping process, Axelrod is mistaken to conclude that the troops on either side of No Man's Land were really following a tit-for-tat policy. For why would the damping process not mitigate the effects of a single *attack* the same way it mitigates a volley of several retaliations?

[42] Binmore notes that several other authors have also given proofs of this result, starting with Selten (1983). Despite this, many continue to believe that tit-for-tat is evolutionarily stable, a part of the "tit-for-tat bubble". Binmore attributes the continuing confusion in part to a proof in Maynard Smith (1982) that tit-for-tat is *weakly* evolutionarily stable. A weakly evolutionarily stable strategy cannot repel all small-scale mutant invasions. So if a strategy has the weak evolutionarily stability property, this implies only that there is some polymorphism in which this strategy can survive as a fraction of the population.

coexisting strategies people employ in their repeated interactions. In support of this conclusion, Binmore points to an important study by Linster (1992). Linster tested the robustness of Axelrod's computer simulation results by running a new series of simulations, in which the sixty-three tournament strategies are matched against each other in an infinitely repeated Prisoners' Dilemma and evolve according to the replicator dynamics. In Axelrod's original simulation of evolution, tit-for-tat emerged as the strategy claiming the largest share of the population because Axelrod chose initial conditions of the simulation especially favorable to tit-for-tat.[43] Linster systematically investigated all possible initial conditions, and after running his series of computer simulations found that tit-for-tat is played most frequently in the final polymorphism only about one-fourth of the time. Linster then ran a second series of simulations which produced further striking results. There is nothing intrinsically special about the sixty-three strategies of Axelrod's tournament. Evolutionary game theory investigates which strategies should survive and which should die out over time, so a proper evolutionary analysis of any game should investigate all possible strategies, or at least, all strategies of a certain complexity.[44] One measure of a strategy's complexity is the number of states required to define an automaton that characterizes this strategy. Twenty-six strategies for playing the repeated Prisoners' Dilemma can be defined with an automaton having at most two states. Linster ran simulations of these twenty-six strategies against each other under a variety of initial conditions. In these simulations, no single strategy ever drives all others to extinction. At the end of each simulation, the grim strategy claims over half the population. A portion of tit-for-tat-followers can also remain in the final polymorphism, as well as a portion of those who cooperate unconditionally, but Linster's

[43] While Binmore (p. 315) is right to observe that Axelrod set the initial conditions of his simulation in a way especially favorable to his own conclusion, there is no reason to suppose that Axelrod did so intentionally.

[44] In *Fun and Games*, Binmore (1992, p. 353) informs the reader that in a repeated Prisoners' Dilemma with just 10 possible repetitions, each agent has $2^{349,525}$ pure strategies. So computer simulations of repeated Prisoners' Dilemma which incorporate all possible strategies seem to be out of our reach.

analysis supports the conclusion that grim-followers will tend to evolve more readily than tit-for-tat followers. Moreover, Binmore cites a recent study by Probst (1996), who ran computer simulations of repeated Prisoners' Dilemma with a more complex evolutionary dynamics that permits mutant strategies characterized by automata of up to twenty-five states to enter into the population. In Probst's simulations, strategies that are initially cooperative, like tit-for-tat and grim, tend to be overwhelmed by strategies that start with defection and only switch to cooperation when matched against strategies that will retaliate against defection. So from the evidence from computer simulations, we cannot even conclude that evolution favors initially "nice" strategies like tit-for-tat or grim.

Finally, Binmore considers a question which turns out to be crucial for political philosophy: Who enforces a pattern of social cooperation in settings prone to free-riding? The "tit-for-tat bubble" masks the importance of this question. If we were a society of tit-for-tat-followers, then each one of us would punish another who defected in one encounter in the very next encounter, so one would always enforce cooperation with oneself. However, this kind of answer to the question assumes that all our encounters are pairwise, and that a defector will be punished by the injured party. In fact, opportunities for mutual aid in society need not be pairwise, and in many settings an injured party will not find it desirable or even possible to punish a defector. This gives us reasons beyond the evidence of computer simulations for thinking that human societies are not typically tit-for-tat societies. At the same time, this raises a serious challenge to any equilibrium explanation of social cooperation. If one is dealing with another agent who will not, or cannot, retaliate against a defector, then to defect is apparently the rational course of action. This is the challenge of Hobbes' Foole (1651, Chapter 15), and if the Foole is right, then rational agents will exploit the helpless in society without mercy. Gauthier (1986, p. 17) and Barry (1989, p. 163) regard this as a particularly serious problem for anyone who tries to analyze justice in terms of mutual advantage.

Binmore, who explicitly places his work within this tradition, deals with the problem in a way reminiscent of Hobbes' response to the Foole and Hume's explanation of why people keep their promises (1740, pp. 521-522). Hobbes and Hume both argue that someone who follows the Foole's advice and defects against another who cannot retaliate now is liable to be defected against in the future, if not by the injured party then by others who will not be willing to cooperate with a known exploiter. Binmore, reasoning along similar lines, uses an *overlapping generations model* to argue that agents can sustain an equilibrium of mutual cooperation in an indefinitely repeated mutual aid game even if no agent can ever punish another who exploits her. In Binmore's model, each agent is assumed to have one predecessor and one descendant. Each agent is productive during the first period of his life, which immediately follows the productive period of his ancestor's life. During one's productive period one can either cooperate by sharing the (perishable) product of his efforts with his now unproductive ancestor, or defect by consuming all of this product. Can the agents expect each other to cooperate over time, given that one who is no longer productive cannot retaliate if his descendant defects? Binmore puts his own model to the test by assuming that all are heartless towards their helpless ancestors. So the base game is structurally similar to the Prisoners' Dilemma, in that all agents prefer universal cooperation to universal defection, but to defect is each agent's unique best move. The Foole will recommend that any agent defect during his productive period, since his ancestor cannot punish him. But Binmore points out that there is an equilibrium of mutual cooperation in this repeated game. This cooperative equilibrium is characterized by a *conformist* strategy, which requires one to cooperate if and only if one's ancestor was a conformist. In this equilibrium, cooperative behavior is enforced, not by the exploited, but by the descendants of exploiters! Kandori (1992*a*,1992*b*) proved relevant folk theorems, which show that in a community of overlapping generations, any mutually beneficial outcome of any base game can be sustained as part of an equilibrium of the indefinitely repeated game. Ellison (1994) proved that there is an equilibrium of mutual cooperation in an indefinitely repeated Prisoners' Dilemma even when agents are constantly rematched at random and cannot recognize each other. In Ellison's model, a defector is punished by a stranger, and will tend to revert back to cooperation not to appease his punisher, but to prevent the spread of further defections as a result of his punishment, which might lead to his being defected against even more in the future. From all these results, Binmore (p. 277, pp. 329-334) draws an important philosophical conclusion: To view justice as a system of mutual advantage does not imply that the powerless in society can be exploited with impunity. Indeed, on a justice-as-mutual-advantage view, we may be required to help those who cannot help themselves, so as to encourage those who can to help us.

Binmore presents a powerful case that the story of the evolution of cooperation in human societies is a great deal more complicated than Axelrod's early work suggests. Evolutionary game theory does not support the claim that behavior in human societies converges to a focal norm of cooperation like tit-for-tat or grim, after all. Binmore's work should shift our understanding of social cooperation. When the members of society are cooperating towards the common good rather than free-riding off one another, this is likely to be the result of different segments of society following different norms of cooperation, so that a plurality of cooperative norms coexist and support a system of mutual aid. However, these norms do not necessarily support each other. We live in a world in which different creeds jockey for our attention and our allegiance. Evolutionary game theory models this basic fact of life. Much more work remains to be done before we will come close to fully understanding the evolution of cooperation. Binmore is probably right in his prediction that evolutionary game theorists will find questions of strategic complexity to be of increasing importance as this work continues (p. 323). And he is surely right to argue that cooperation in human cultures must be analyzed in terms of some stability concept different from evolutionary stability. Besides the fact that no strategy in the repeated Prisoners' Dilemma is evolutionarily stable, Maynard Smith and

Price's (1973) definition of evolutionary stability considers only cases in which mutant strategies enter into the population one at a time, and as Binmore notes, in human societies any norm of cooperation is prone to bombardment by several different "rebel" behaviors simultaneously (p. 268, p. 326).

However, Binmore perhaps underrates the role that "nice" individuals might play in the evolution of cooperation. True, many cases of "spontaneous" cooperation, like Axelrod's trench war examples, admit of a "mean" strategy explanation rather than a tit-for-tat explanation. The evidence from computer simulations also indicates that "mean" strategies as well as "nice" strategies can evolve in populations. But even if it turns out to be true that evolution does not *favor* "niceness", it may be the case that a pattern of mutual aid frequently emerges within a population because the *presence* of nice individuals becomes common knowledge and encourages cooperative behavior in others. McKelvey and Palfrey (1992) explored this possibility in a landmark experimental study of the *centipede game*, which is a game of alternating moves in which to cooperative is to increase the common good and give the other agent the opportunity to cooperate or to free-ride, and to free-ride is to claim most of the current common good and end the game. To free-ride is each agent's unique best move at every stage of the centipede, but McKelvey and Palfrey found that pools of experimental subjects cooperate far more often than they "should" in centipede games. They were able to account for this behavior by assuming that it was common knowledge among subjects that some small proportion of them are "nice" and will cooperate throughout.[45] What spurred cooperation in a set of

[45] In McKelvey and Palfrey's experiment, pools of twenty subjects were matched to play either a 4- or a 6-move centipede game with monetary payoffs. Each subject participated in nine or ten centipede games, each with a different partner. Of 662 games played in the study, only 37 (5.6%) ended with the first agent defecting on the first move, even though this outcome characterizes the unique Nash equilibrium in terms of the monetary payoffs. McKelvey and Palfrey found that approximately 5% of the experimental subjects were "nice" individuals who cooperate throughout. They attribute the widespread cooperation in their experiment to the "selfish" subjects cooperating early in the centipede to test the other subject, in hopes that one has met a "nice" person who will cooperate and thus increase the "selfish" agent's payoff.

McKelvey and Palfrey's study has spurred a controversy over *why* the "nice" individuals cooperate throughout in the centipede. McKelvey and Palfrey originally conjectured that the "nice" subjects were to some extent altruistic, in the sense that they preferred to have others receive a larger share of the common experiments with centipede games might also spur the evolution of cooperation in a larger society. The role that "nice" individuals might play in evolutionary game theory is a subject for continuing study.[46]

§4. The Social Contract as Equilibrium

> we are not bound together by iron shackles of duty and obligation, but by common understandings and conventions that propagate through societies like fungus through a rotten log. Each single filament in the system is so fragile that it can hardly survive the light of day. But, like Gulliver in Lilliput, we are no less tightly bound by the totality of such gossamer threads than if Nature had really provided us with the shackles that traditionalists invent. (*Just Playing*, p. 226)

Evolution of the Social Contract stresses the value of evolutionary game theory as a tool for explaining social norms, ranging from norms of fair distribution, norms governing personal property; norms which require individuals to help each other produce a common good and even norms of language. The theory explains not only how a society sustains all of these norms, but how they originate. True, the explanations are in rough outline form, but we can expect them to become more precise as evolutionary game theory continues to develop. Nevertheless, we must consider an objection to Skyrms' and Binmore's analytical framework, namely, that this framework seems to leave no room for normative ethics. The final state evolutionists fallaciously conclude that we ought to create the society that evolution will eventually produce. As we have seen, on Skyrms' and Binmore's view, evolution can produce a number of different societies, which are

good over exploiting others. In private correspondence, McKelvey and Palfrey informed me that they now believe it possible that the "nice" individuals in the study were following a fairness norm similar to the fairness norm that seems to influence behavior in ultimatum game experiments.

[46] Almost needless to say, the role that "nasty" individuals play in evolutionary game theory needs to be studied further, as well. One might plausibly suppose that a "nasty" individual might encourage uncooperative conduct in others, just as a "nice" individual might encourage others to act cooperatively.

characterized in part by their distinctive systems of norms. But if the various systems of social norms in the world really are the product of cultural evolution, is it unnecessary, even fallacious, to argue that we should strive to lead better lives? Is it pointless to argue that we should sometimes reform our behavior and our institutions?

Binmore and Skyrms both anticipate this objection. Their response is that evolutionary game theory provides a framework for explaining how current social norms can be changed as well as how they can persist, and that this opens the possibility of a game-theoretic account of social reform. On their view, a society's social contract is the set of norms which guide the conduct of its members. A social contract is therefore an equilibrium of very complex game, the game of all of the interactions of society. A society can reform itself by moving from its current equilibrium to a new social contract. Skyrms gives no specific recommendations for social reform, but closes *Evolution of the Social Contract* by pointing out that "Even those who aim to change the world had better first learn how to describe it. (p. 109)". A prerequisite to reforming the social contract is understanding the nature of this contract, and as Skyrms sees it, evolutionary game theory is the vehicle by which we can reach this understanding.

Binmore takes up the question of reforming the social contract in a highly original way. Binmore himself describes his approach as a synthesis of Harsanyi's and Rawls' theories of distributive justice (1994, p. 52). One might also describe Binmore's approach as an extension of evolutionary game theory with a variation of Rawls' and Harsanyi's device of the original position. Harsanyi and Rawls argue that one should accept certain principles as principles of justice because these are the principles that rational agents would choose from an original position in which they know nothing that would enable them to choose principles which favor anyone in particular. Rational choice behind such a veil of ignorance is the cornerstone idea of justice as fairness. Ever since Harsanyi and Rawls introduced this idea, critics have claimed that justice as fairness commits actual persons in the world to nothing because the choice of principles is merely hypothetical.

Rawls (1971, pp. 19-21) argues in reply that the principles chosen in the original position do have moral force because this kind of choice embodies certain widely held views, in particular that principles of justice should be rational and impartial and that they should cohere with certain considered moral judgments, such as the judgment that slavery and serfdom are immoral institutions. Harsanyi's response is similar. In Harsanyi's theory, the original position is simply a device that ensures that *a priori* the interests of all members of society will be weighted equally in determining principles of justice. This reflects a fundamental principle, namely, that society should disadvantage no one arbitrarily (Harsanyi 1975). Harsanyi and Rawls are prepared to defend their alternative versions of justice as fairness by appealing to certain very general moral principles. But what if one does not accept these principles? If justice as fairness is supposed to vindicate a particular system of justice, then one might argue that Rawls and Harsanyi ought not appeal to any moral claims in defense of the moral claims chosen in the original position. Moreover, even if one admits that the principles chosen in the original position have a certain moral force, this does not explain what would enforce these principles in actual society. An agreement, even an *actual* agreement, that certain principles are indeed principles of justice does not by itself commit anyone to obey any of these principles.[47]

The difficulties with justice as fairness stem in large part from the view that final principles of justice can be established outside of time and circumstances. Harsanyi and Rawls both assume that parties choose the principles of justice within the original position at no particular time, and without reference to any particular *status quo* point in

[47] Harsanyi (1980, p. 127) explicitly acknowledges this, and thus unlike Rawls does not want his version of justice as fairness to be thought of as a contractarian theory. Harsanyi argues that any moral code is rationally justified because it maximizes expected social utility, not merely because any group of individuals may have agreed to follow this code. To explain why individuals would be rational to *follow* the requirements of a rule-utilitarian code, such as a requirement to vote, Harsanyi argues that one can ask the question "What would happen if people like me did not vote?" and conclude that a *commitment* to vote is rational if social utility is increased when all like oneself vote (p. 130). Binmore rejects this kind of Kantian argument outright.

which society would find itself if there is no agreement in the original position. This has led to charges that justice as fairness ignores a society's history (Wolff 1966, Gellner 1988). Additionally, justice as fairness prescribes a certain kind of society without explaining how this society is to be achieved. Rawls argues that the original position parties will select as the principle of distributive justice a *difference principle*, which requires that primary social goods be distributed to the greatest advantage of the least well-off in society. The difference principle is a special case of the *maximin* rule, which recommends choosing an option one most prefers assuming that one's worst-case scenario given one's choice will occur. Rawls spends a considerable part of the latter half of *A Theory of Justice* arguing that his version of justice as fairness will if implemented result in a stable society. This might be so, but there is good reason to doubt that Rawls' principles of justice would be chosen in the original position. Harsanyi (1975) argues that original position parties would not select a principle for distributive justice based upon the maximin rule, which is not only a pessimistic but in certain ordinary contexts a highly irrational principle of choice. Harsanyi (1955, 1977) contends that behind the veil of ignorance, agents should follow the principle of Bayesian rationality and choose principles of justice that maximize the average expected utility for the members of society. Binmore concurs with Rawls that a more egalitarian society produced by following the difference principle will be self-sustaining, *if* society has already reached an equilibrium in which the difference principle is the norm of distributive justice. But Binmore also concurs with Harsanyi that the result of the original position choice will be a form of utilitarianism, not the difference principle. However, this result can only be realized in a society of individuals capable of following what Harsanyi (1955) calls their *ethical preferences* rather than their *subjective preferences*. In agreement with many others, Binmore argues that utilitarianism would demand greater sacrifices from individuals than they would be prepared to make voluntarily (pp. 259-261). In the end, Binmore thinks that neither Rawls nor Harsanyi provide good grounds for thinking that people will comply with the principles of justice chosen in the original position. Rawls and Harsanyi can argue that one has a certain moral obligation to follow the principles of justice as fairness, but on Binmore's view this is simply to say "one ought to do something because one ought to do it (p. 230)". Binmore concludes that without some enforcement mechanism external to the society to be governed according to the original position choice, one cannot expect the members of this society to conform their behavior with this choice, and by construction justice as fairness rules out such mechanisms.

Binmore proposes to apply a somewhat different version of the original position in a novel way, which avoids the need for any robust moral premises and at the same time deals with the commitment problem. Binmore calls his own theory a theory of *the seemly*, to be contrasted with theories of *the good* and *the right* (p. 245). One way to understand this distinction is that in contrast with philosophers who take either consequentialist or deontological notions to be fundamental in moral theory, Binmore makes *equilibrium* the fundamental notion in his moral theory. Binmore even argues that one can give an account of the good as derivative of principles that have evolved for equilibrium selection (Chapter 2), and an account of the right as derivative of rules that have evolved which help to sustain an existing equilibrium (Chapter 3). On Binmore's account of the social contract there is no commitment problem. The constituent norms of a social contract are the "gossamer threads" that bind the members of society to act in certain ways. A reasoned reform[48] of the current social contract is possible, but only if this is achieved by a series of moves from the current social contract to a new set of norms constituting a new social contract, each of which is itself an equilibrium in the complex *"game of life"*. To achieve such a reform, Binmore proposes extending the "game of life" into a *"game of morals"* in which each agent has the right to appeal at any time to a version of the original position, under which all will be behind a veil of

[48] As opposed to the abrupt change in a social contract that might occur after a violent revolution or conquest.

ignorance so that none will know which particular role she plays in the game of life (Chapter 4). Binmore does not justify using a version of the original position in his theory on moral grounds. Instead, he argues that in human cultures, certain mechanisms have evolved for selecting equilibria in bargaining problems that all commonly regard as fair, and that these mechanisms closely resemble the device of the original position. A classic example is the chocolate cake problem, in that people typically consider a division of the cake fair if the one who divides the cake does not know in advance which piece he gets. Binmore sees himself giving a purely pragmatic argument: If the original position comes close to describing how we solve certain bargaining problems in real life, why not use this device to select a new social contract (p. 8, Chapter 2)? Like the original position in Harsanyi's and Rawls' justice as fairness, Binmore's original position has the effect of allowing the members of society to decide upon a new social contract while each believes her new place in society will be chosen at random. However, in Binmore's theory the *status quo* point is made explicit, namely, it is society's current social contract. Also, a choice made in the original position in Binmore's game of morals need not be final. To illustrate this point, let us return to Braithwaite's resource division problem. If Matthew and Luke find themselves with a windfall of some resource, and their payoffs are characterized by the Figure 2 game, then they could divide the resource according to the decision they would make in the original position. Suppose that in the original position, in which they forget which role each plays in the game, Luke and Matthew agree to follow a correlated equilibrium at which they follow (D, H) if a tossed fair coin comes up "heads" and (H, D) if this coin comes up "tails". Upon exiting the original position the coin comes up "tails". Matthew is disappointed, but if one assumes that the agreement in the original position is binding, then Matthew must live with this inequitable outcome. Of course, in such a situation we might plausibly expect Matthew to plead for a new agreement, perhaps a correlated equilibrium defined by a new coin toss. In Binmore's theory, dissatisfied agents can ask for a return to the original position to renegotiate the social contract, knowing that anyone else can do the same. Under these circumstances, the point which Matthew and Luke will settle upon is the fair division point, since any arrangement that leaves either with less than the other will result in an immediate call for a new round of negotiations behind the original position. More generally, if conditions have developed so that the current social contract is no longer optimal, then Binmore argues that a new social contract can be established via the game of morals, but the agents will only regard as viable candidates for a new contract those outcomes which are renegotiation-proof, in the sense that from these outcomes no one will want to call for yet another round of negotiations in the original position. Remarkably, these outcomes coincide with the outcomes that Rawls argues the original position parties would settle upon by using the maximin rule. As Binmore puts it, "a maximin rabbit has therefore been extracted from a Bayesian hat (p. 437)."

So Binmore has defended what is in essence a Rawlsian system of distributive justice with an argument quite different from the sorts of arguments Rawls uses to defend the difference principle. Binmore's approach has certain distinct advantages over that of Rawls. Binmore does not ignore a society's history the way that Harsanyi and Rawls seem to. Indeed, Binmore takes all of a society's history into account, since the *status quo* point is the actual social contract that cultural evolution has produced. Binmore's argument relies only on the idea that a new social contract should be renegotiation-proof. In his theory, there is no need to defend his use of the original position on moral grounds. Unlike Harsanyi and Rawls, Binmore does not assume that the original position reflects any general moral principles at all. Rather, Binmore views the original position as an outgrowth of processes which have evolved for purposes of dividing certain resources (Chapter 2). Moreover, by establishing the *status quo* point of his theory as the *status quo* of actual society, and by requiring that the transition to the new social contract chosen in the game of morals be a continuous path of equilibria joining the old and the new social contracts (Chapter 4), Binmore's theory deals with the problem of commitment to the reformed social contract quite smoothly. Since society is at equilibrium throughout the transition from old social contract to new, each member of society will have a good prudential reason to follow each stage of the transition, which culminates in the new social contract. So Binmore avoids the irrational commitments he thinks Rawls' and Harsanyi's theories would require of individuals in actual society. One might say that Binmore incorporates a game-theoretic version of the "ought-implies-can" principle into his theory.

Binmore's daring idea is that one can develop a moral theory without taking either a conception of the right or a conception of the good as fundamental. In Binmore's theory, both the good and the right are relativized to a particular social contract, that is, to a particular equilibrium of the "game of life". People regard a state of affairs as good because such a state would be picked out by the fairness norms which have evolved in human societies (Chapter 2). To do one's duty is to follow one's part of the social contract. One has a right to take an action if no norm in the social contract prohibits this action. One has a right to something if some norm in the social contract requires others to provide it (Chapter 3). Binmore candidly endorses a moral relativism in which there is no good or right in any absolute sense. Indeed, as Binmore develops his case in *Just Playing*, he argues vigorously against those who defend either absolute goods or absolute rights. In particular, Binmore argues that it is a mistake to think that any social practice or institution is unjust in principle. Binmore's overlapping generations example shows that there can be social contracts in which the powerless in society have positive rights, but nothing guarantees that a particular social contract will either protect the powerless from exploitation or provide for any of their needs. And Binmore takes the fact that many actual social contracts seem to grant few or no rights to the powerless as evidence that his conception of moral theory reflects actual moral practice far better than theories which appeal to some absolutist conception of the good or the right.

> The unwelcome truth is that practical morality --- the morality by which we actually live --- does in fact endorse the exploitation of those powerless to resist. We dismiss the homeless and the destitute as being an unfortunate consequence of the necessity that a productive society provide adequate incentive for its workers. Is this not to accept that an underclass must suffer in order that the rest of us can enjoy a higher standard of living? We do not, of course, say this openly. Instead, we square things with our consciences by dehumanizing those excluded from the feast. (pp. 258-259)

Binmore is sure to receive criticism for his controversial claims about the right and the good as spirited as his own criticisms of universalistic or absolutist accounts of the good and the right as *ipsedixism* (p. 152).[49] Here I will not evaluate Binmore's arguments for these claims. Binmore's theory raises what I think are two more immediate questions: (1) Must so-called moral "ipsedixists" reject Binmore's general methodology for reforming the social contract?, and (2) Does this methodology vindicate "maximin principle justice"? So far as I can see, one need not be committed to the moral relativism to which Binmore commits himself in order to use his game of morals to select a new equilibrium in the game of life. One could, for instance, take a universalistic account of the good as fundamental in one's moral theory, and then view the various social contracts that could emerge according to Binmore's theory as alternative ways in which the members of a community can best achieve the good. On the other hand, I suspect that some philosophers and social scientists will have serious objections to Binmore's naturalistic defense of Rawls' difference principle as a principle of distributive justice. In *Just Playing*, Binmore shows analytically that the social contract that agents choose in the game of morals conforms with the difference principle. So Binmore's critics will have to

[49] As Binmore acknowledges, Bentham (1789) coined the term 'ipsedixist' to refer to one who cloaks one's own prejudices as moral imperatives all should follow.

challenge the assumptions of Binmore's model. Perhaps somewhat surprisingly, I suspect that the main objection to Binmore's model will turn out to be his assumption that the equilibrium that agents select in the game of morals be renegotiation-proof. In Binmore's game of morals, if the agents were to choose any social contract in the original position that did not satisfy the difference principle, then the least advantaged would immediately call for a reentry into the original position. But suppose the members of society can make a transition from the *status quo* to a new social contract which guarantees that everyone receives what all agree is an acceptable minimum payoff and maximizes the average expected payoff subject to this guarantee, so that one's expected payoff is significantly greater than one's expected payoff in a social contract which conforms to the difference principle. If this social contract leaves the least well-off with even slightly less than they would receive according to the difference principle, then in Binmore's model the least well-off will reject it. Yet one could argue that the rejected social contract is really more viable than the maximin social contract, since nearly everyone has a good reason to complain that the maximin contract deprives them of significant gains they would have achieved under the rejected contract, while the least well-off would still have had an acceptable minimum.[50] This argument echoes some of the criticisms that Harsanyi (1975) and Kavka (1986) raised against the difference principle in Rawls' theory. Some of Binmore's critics may take this argument to imply that the assumptions in Binmore's model should be relaxed, so that a few holdouts cannot block society from achieving certain social contracts. Nevertheless, while some may disagree with certain details in Binmore's account of the game of morals, he has provided us with an ingenious new approach to arguing for alternative social contracts.

[50] In a striking experiment designed by Frohlich, Oppenheimer and Eavey (1987), groups of five students were assigned the problem of agreeing upon a rule for distributive justice subject to the constraint that no one knew her future position in the distribution. Of the 44 groups that reached unanimous agreement, 35 (79.5%) chose a distribution which maximizes the average subject to a floor constraint, and *none* chose the difference principle!

§5. Conclusion

As I hope the sampling of topics from their work discussed above illustrates, the idea of applying evolutionary game theory in moral philosophy has come of age with Binmore's and Skyrms' works. Jointly, these works make by a wide margin the most significant step so far in the general research program Braithwaite began with his 1954 Cambridge lecture. Together, *Evolution of the Social Contract* and *Just Playing* deserve to make an impact comparable to that Alasdair MacIntyre made with the publication of *After Virtue* in 1981. It is now a commonplace that MacIntyre stimulated a new interest in the role of the virtues in moral theory in the 1980s. Skyrms' and Binmore's books should at the end of the century spark at least as much interest in evolutionary game theory as an analytical framework for moral philosophy.

Interestingly, *Just Playing* shares both the strengths and the weaknesses of *After Virtue*. Binmore's work reflects an amazing learning. His bibliography alone is worth the price of the book. *Just Playing* has the same intellectual sweep and is as exciting to read as *After Virtue*. And like *After Virtue*, *Just Playing* is filled with interesting criticisms of contemporary moral theory and presents a powerful argument for a neglected alternative, in this case a moral theory based upon equilibrium points of the game of life. On the negative side, Binmore frequently misrepresents the views of other authors, as when he claims that Rawls tried to provide a "moral geometry" in *A Theory of Justice* and that Rawls presents justice as fairness as a theory which deduces the good from the right. Binmore is also sometimes less than fair to positions he opposes. For instance, it is not so clear as Binmore claims that utilitarianism demands too much of individuals. Harsanyi (1980, 1992) explicitly argues that a system of rule utilitarianism would include rights that would protect individuals from having to make excessive sacrifices for the common good. To his credit, Binmore does treat both Harsanyi's case for utilitarianism and Rawls' case for the difference principle seriously. Elsewhere in *Just Playing*, Binmore's dismissals of certain important views in moral philosophy are sometimes so curt and derisive that I am left wondering if Binmore is trying to outrage his readers in hopes of stimulating their interest in his own views --- yet again, I perceive a similarity between Binmore and MacIntyre. Finally, *Just Playing* takes numerous detours into topics ranging from proposals for generalizing the Nash bargaining model to the free will-determinism debate. Binmore's discussions of these tangential topics are lively and thought provoking, as is nearly every paragraph of *Just Playing*, but they unfortunately tend to obscure his central arguments. Readers may find it hard to follow the main thread of argument in *Just Playing*, even with the aid of both a Reading Guide which Binmore provides (*xix-xxiii*) and the opening chapter of *Playing Fair*, which Binmore wrote to provide an overview of his theory. Overall, the flaws of *Just Playing* are distracting, but they do not diminish the significance of Binmore's contributions to moral philosophy. Binmore's positive theory of distributive justice rivals those of Rawls and Harsanyi in importance. And while his criticisms of other moral philosophers may not always be on the mark, Binmore has given us the most sophisticated defense of a convention-based moral relativism we are likely to get. Readers will have to study *Just Playing* with some patience, but they will be more than amply rewarded for their efforts.

Evolution of the Social Contract is almost an opposite extreme from *Just Playing* in style, though certainly not in the quality of its content. Skyrms' discussion is characteristically no-nonsense, elegant and penetrating, sometimes almost deceptively so. Skyrms presents his original arguments in *Evolution of the Social Contract* with so little fanfare that readers may not always immediately appreciate their significance. A case in point is what may look like Skyrms' almost effortless introduction of correlated interactions in the fair division problem and the one-shot Prisoners' Dilemma, which actually generalizes evolutionary game theory into a far more predictively powerful theory. If readers have any disappointment with *Evolution of the Social Contract*, it will likely be that they will sometimes wish Skyrms had said more. As noted several times in the above sections, the ideas Skyrms presents in *Evolution of the Social Contract* can be extended into a number of research programs. Skyrms does not pursue all of these various programs in *Evolution of the Social Contract.* As Skyrms says in his Postscript (p. 105), he does not attempt to present a complete theory of the social contract. Skyrms' aims are purely explanatory. Some readers might want to see what sort of normative conclusions Skyrms would draw from his evolutionary framework.

As with any important philosophical works, both *Evolution of the Social Contract* and *Just Playing* will launch a critical literature that will further develop the authors' insights, sometimes in directions neither would have predicted.[51] There will surely even be unexpected philosophical fruit from authors who will disagree with large parts of Skyrms' and Binmore's overall position.[52] However the discussion they have started proceeds, Binmore and Skyrms have secured a permanent place for evolutionary game theory in moral philosophy.

REFERENCES

Anscombe, G. Elizabeth. 1958. 'Modern moral philosophy.' *Philosophy*, 33: 1-19.

Aumann, Robert. 1974. 'Subjectivity and correlation in randomized strategies.' *Journal of Mathematical Economics*, 1: 67-96.

Aumann, Robert. 1987. 'Correlated equilibrium as an expression of Bayesian rationality.' *Econometrica*, 55: 1-18.

Axelrod, Robert. 1981. 'The emergence of cooperation among egoists'. *American Political Science Review*, 75, 306-318.

Axelrod, Robert. 1984. *The Evolution of Cooperation.* New York: Basic Books, Inc.

[51] This is already happening in Skyrms' case. As noted above (note 23), D'Arms, Batterman and Gorny are exploring the properties of anti-correlated strategies in the bargaining problem. Also, several authors will be commenting on *Evolution and the Social Contract* in a forthcoming book symposium issue of *Philosophy and Phenomenological Research.*

[52] Something like this has happened as a result of the virtue-ethics movement launched by Anscombe (1958) and MacIntyre. For instance, Anscombe's and MacIntyre's sometimes disparaging appraisals of Hume, Kant and Sidgwick for their alleged neglect of the virtues has encouraged new research into the important roles the virtues actually play in Hume's, Kant's and Sidgwick's moral theories.

Barry, Brian. 1989. *Theories of Justice*. Berkeley: University of California Press.

Bentham, Jeremy. (1789) 1970. *An Introduction to the Principles of Morals and Legislation*. Darien, Connecticut: Hafner Publishing Company.

Bicchieri, Cristina. 1993. *Rationality and Coordination*. Cambridge: Cambridge University Press.

Binmore, Ken. 1987*a*. 'Modeling rational players I'. *Economics and Philosophy*, 3: 9-55.

Binmore, Ken. 1987*b*. 'Modeling rational players II'. *Economics and Philosophy*, 4, 179-214.

Binmore, Ken. 1992. *Fun and Games: A Text on Game Theory*. Lexington, Massachusetts: D. C. Heath and Company.

Binmore, Ken. 1994. *Game Theory and the Social Contract Volume I: Playing Fair*. Cambridge, Massachusetts: MIT Press.

Binmore, Ken. 1998. *Game Theory and the Social Contract Volume II: Just Playing*. Cambridge, Massachusetts: MIT Press.

Binmore, Ken and Samuelson, Larry. 1992. 'Evolutionary stability in repeated games played by finite automata.' *Journal of Economic Theory*, 57, 278-305.

Binmore, Ken, Gale, John and Samuelson, Larry. 1995. 'Learning to be imperfect: The ultimatum game'. *Games and Economic Behavior*, 8, 56-90.

Binmore, Ken and Samuelson, Larry. 1996. 'Muddling through: Noisy equilibrium selection'. (to appear in *Journal of Economic Theory*).

Braithwaite, Richard. 1955. *Theory of Games as a Tool for the Moral Philosopher*. Cambridge: Cambridge University Press.

Brandenburger Adam and Dekel, Eddie. 1988. 'The role of common knowledge assumptions in game theory', in *The Economics of Missing Markets, Information and Games*, ed. Frank Hahn. Oxford: The Clarendon Press. 46-61.

Crawford, Vincent. 1997. 'Theory and experiment in the analysis of strategic interaction', in *Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress, Vol. I,* Econometric Society Monographs No. 27, ed. David Kreps and Ken Wallis. Cambridge, U.K., and New York: Cambridge University Press: 206-242.

D'Arms, Justin. 1996. 'Sex, fairness, and the theory of games'. *Journal of Philosophy,* 93, 615-627.

D'Arms, Justin, Batterman, Robert and Gorny, Krzysztof. 1997. 'Game theoretic explanations and the evolution of justice'. (to appear in *Philosophy of Science*).

Ellison, Glenn. 1994. 'Cooperation in the Prisoner's Dilemma with anonymous matching.' *Review of Economic Studies,* 61: 567-588.

Foster, Dean and Young, H. Peyton. 1990. 'Stochastic evolutionary dynamics.' *Journal of Theoretical Biology,* 38, 219-232.

Frohlich, N., Oppenheimer, J. A. and Eavey, C. L. 1987. 'Laboratory results on Rawls' distributive justice'. *British Journal of Political Science,* 17, 1-21.

Fudenberg, Drew and Levine, David. 1998. *The Theory of Learning in Games.* Cambridge, Masachusetts: MIT Press.

Gauthier, David. 1986. *Morals By Agreement.* Oxford: Clarendon Press.

Gellner, Ernest. 1988. *Plough, Sword and Book.* London: Paladin, Grafton Books.

Guth, Werner, Schmittberger, Rolf and Schwarze, Bernd. 1982. 'An experimental analysis of ultimatum bargaining'. *Journal of Economic Behavior and Organization,* 3, 367-388.

Hampton, Jean. 1986. *Hobbes and the Social Contract Tradition.* Cambridge: Cambridge University Press.

Hampton, Jean. 1987. 'Free riders and collective goods'. *Economics and Philosophy,* 3, 245-273.

Harsanyi, John. 1955. 'Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility'. *Journal of Political Economy*, 63, 309-321.

Harsanyi, John. 1975. 'Can the maximin principle serve as a basis for morality? A critique of John Rawl's theory'. *American Political Science Review*, 59, 594-606.

Harsanyi, John. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Harsanyi, John. 1980. 'Rule utilitarianism, rights, obligations and the theory of rational behavior.' *Theory and Decision*, 12, 115-33.

Harsanyi, John. 1992. 'Game and decision theoretic models in ethics' in *Handbook of Game Theory*, Vol. 1, ed. Robert Aumann and Sergiu Hart. Elsevier Science Publishers B. V. 669-707.

Hobbes, Thomas. (1651) 1991. *Leviathan*, ed. Richard Tuck. Cambridge: Cambridge University Press.

Hume, David. 1740 (1888, 1976). *A Treatise of Human Nature*. ed. L. A. Selby-Bigge. rev. 2nd. ed. P. H. Nidditch. Oxford: Clarendon Press.

Hume, David. 1777 (1888, 1975). *An Enquiry Concerning the Principles of Morals*, ed. L. A. Selby-Bigge. rev. 3rd. ed., ed. P. H. Nidditch. Oxford: Clarendon Press.

Huxley, Thomas H. 1888. 'The struggle for existence and its bearing upon man.' *Nineteenth Century* 23: 161-180.

Kandori, Michihiro. 1992*a*. 'Social norms and community enforcement.' *Review of Economic Studies*, 59: 63-80.

Kandori, Michihiro. 1992*b*. 'Repeated games played by overlapping generations of players.' *Review of Economic Studies*, 59: 81-92.

Kalai, Ehud and Smorodinsky, Meir. 1975. 'Other Solutions to Nash's Bargaining Problem'. *Econometrica* 16: 29-56.

Kavka, Gregory. 1986. *Hobbesian Moral and Political Theory*. Princeton: Princeton University Press.

Lewis, David. 1969. *Convention: A Philosophical Study.* Cambridge, Massachusetts: Harvard University Press.

Linster, Bruce. 1992. 'Evolutionary stability in the infinitely repeated Prisoners' Dilemma played by two-state Moore machines.' *Southern Economic Journal*, 58: 880-903.

Luce, R. Duncan and Raiffa, Howard. 1957. *Games and Decisions: Introduction and Critical Survey.* New York: John Wiley and Sons.

Mackie, John. 1978. 'The law of the jungle: Moral alternatives and principles of evolution'. *Philosophy*, 53: 455-464.

MacIntyre, Alasdair. 1981. *After Virtue: A Study in Moral Theory.* Notre Dame, Indiana: University of Notre Dame Press.

Maynard Smith, John. 1982. *Evolution and the Theory of Games.* Cambridge: Cambridge University Press.

Maynard Smith, John, and Price, G. R. 1973. 'The logic of animal conflict'. *Nature* 146: 15-18.

McKelvey, Richard and Palfrey, Thomas. 1992. 'An experimental study of the centipede game.' *Econometrica*, 60: 803-836.

Nash, John. 1950*a*. 'Equilibrium points in n-person games.' *Proceedings of the National Academy of Sciences of the United States* 36: 48-49.

Nash, John. 1950*b*. 'The Bargaining Problem.' *Econometrica* 18: 155-162.

Nash, John. 1951. 'Non-Cooperative Games.' *Annals of Mathematics* 54: 286-295.

Nash, John. 1953. 'Two-Person Cooperative Games.' *Econometrica* 21: 128-140.

Nash, John. 1996. *Essays on Game Theory.* Cheltenham, United Kingdom: Edward Elgar.

Nydegger, Rudy V. and Owen, Guillermo. 1975. 'Two person bargaining: An experimental test of the Nash axioms'. *International Journal of Game Theory* 3,239-349.

Ochs, Jack and Roth, Alvin. 1989. 'An experimental study of sequential bargaining'. *American Economic Review* 79, 355-384.

Quine, W. V. 1936 (1983). 'Truth by convention', in *Philosophy of Mathematics: Selected Readings*, 2nd ed., ed. Paul Benacerraf and Hilary Putnam. Cambridge: Cambridge University Press. 329-354.

Probst, D. 1996. *On Evolution and Learning in Games*. Ph.D. thesis, University of Bonn.

Raiffa, Howard. 1953. 'Arbitration schemes for generalized two-person games', in *Contributions to the Theory of Games*, vol. 2. ed. H. Kuhn and A. W. Tucker. Princeton: Annals of Mathematics Studies, no. 28. 361-387.

Ramsey, Frank. 1926 (1931). 'Truth and probability', in *The Foundations of Mathematics and Other Essays*, ed. R. B. Braithwaite. New York: Harcourt Brace. 156-198.

Rawls, John. 1957. 'Justice as fairness.' *Journal of Philosophy*, 54, 653-662.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, Massachusetts: Harvard University Press.

Roth, Alvin. 1995. 'Bargaining experiments', in *Handbook of Experimental Economics*, ed. John Kagel and Alvin Roth. Princeton: Princeton University Press. 253-348.

Russell, Bertrand. 1921. *The Analysis of Mind*. London: George Allen and Unwin Ltd.

Savage, L. J. 1954. *The Foundations of Statistics*. New York: John Wiley and Sons.

Schiffer, Stephen. 1972. *Meaning*. Oxford: Oxford University Press.

Schelling, Thomas. 1960. *The Strategy of Conflict*. Cambridge, Massachusetts: Harvard University Press.

Selten, Reinhard. 1983. 'Evolutionary stability in extensive 2-person games.' *Mathematical Social Sciences*, 5: 269-283.

Sidgwick, Henry. 1902. *Lectures on the Ethics of T. H. Green, Mr. Herbert Spencer and J. Martineau*. London: MacMillan and Co., Ltd.

Skyrms, B. 1990. *The Dynamics of Rational Deliberation.* Cambridge, Massachusetts: Harvard University Press.

Skyrms, Brian. 1991. 'Inductive deliberation, admissible acts, and perfect equilibrium', in *Foundation of Decision Theory*, ed. M. Bacharach and S. Hurley. Oxford and Cambridge, Massachusetts: Blackwell: 220-241.

Skyrms, Brian. 1996. *Evolution of the Social Contract.* Cambridge: Cambridge University Press.

Skyrms, Brian. 1998. 'The shadow of the future', in *Rational Commitment and SOcial Justice: Essays for Gregory Kavka*, ed. Jules Coleman and Christopher Morris. Cambridge: Cambridge University Press: 12-22.

Sugden, Robert. 1986. *The Economics of Rights, Co-operation and Welfare.* Oxford: Basil Blackwell, Inc.

Taylor, Michael and Ward, Hugh. 1982. 'Chickens, whales and lumpy goods: Alternative models of public goods provision.' *Political Science*, 30, 350-370.

Walzer, Michael. 1983. *Spheres of Justice.* New York: Basic Books.

Wolff, Robert Paul. 1966. 'A refutation of Rawls' theorem on justice.' *Journal of Philosophy*, 63: 170-190.

Young, H. Peyton. 1993. 'An evolutionary model of bargaining.' Journal of Economic Theory, 59, 145-168.