Reflections on Ethics and Game Theory

Steven T. Kuhn
Department of Philosophy
Georgetown University
Washington, DC 20015

Telephone: 202 966-1829
Fax:202 687-4493
Email: kuhns@georgetown.edu

Reflections on Ethics and Game Theory

1. Introduction

It is plausible to suppose that morality, at its core, is about meshing our own interests with those of others, sometimes consonant with ours and sometimes conflicting. What we investigate as moral theorists may include (explicit and implicit) rules of conduct, qualities of people and institutions, and attitudes toward and emotions elicited by conduct, people and institutions. The particular rules, qualities, attitudes and emotions that are of interest are ones that restrain us from thwarting the interests of others to better serve our own or which move us to advance mutual interests or to advance the interests of others at the expense of our own. The best developed theoretical framework within which questions about conflict and congruence of interests is addressed is the mathematical theory of games. It is natural, therefore, to expect that game theory might be useful in moral theorizing, that it might serve, as R Braithwaite wrote during its early years of development as "a tool for the moral philosopher." Indeed, one might expect crisp game-theoretic analyses would eventually supplant much of the murkier ordinary language debate that fill the ethics journals. Yet, in spite of a few promising forays by people who might be regarded as insiders to moral philosophy[1] and several instructive suggestions by outsiders,[2] this expectation has not materialized.

Two obstacles, I believe, have thwarted a fuller realization of Braithwaite's vision. First, we have not been sufficiently careful in thinking about the game theoretic framework most appropriate for these discussions and the proper interpretation of the technical devices employed. These may well differ from the frameworks and interpretations suitable for other applications of game theory. Second, a number of examples from game theory seem to challenge its purported relevance to moral philosophy, or at least to raise questions about the proper relations between the two disciplines. In this paper I will first discuss several important foundational issues. Then I will present four of the challenging examples and begin to speculate about what lessons we ought to draw from them.

I am afraid that much of what follows will seem (correctly) to be unsolicited exhortations to others. One of the most influential papers in philosophical logic over the last few decades began with the observation that it is much more pleasant to give advice to receive it. We on the less pleasant side of the subsequent transactions were happy to accept our role because of the author's eminence and talents. I have no illusions that my own advice will be so cheerfully accepted. I do hope, however, that it might inspire some further examination of the connections between ethics and game theory.

2. The Framework

2.1. Iteration

The action in game theory seems to have shifted from the one-shot game to the repeated game and especially to evolutionary versions of such games in which the more successful strategies replicate themselves and the less successful die out. Those of us interested in moral philosophy should be a little careful about joining the bandwagon. It is true that many actions that concern moral philosophers--paying debts, helping others, honoring promises, keeping confidences--are the kinds of actions that *are* performed repeatedly and it is not unreasonable to suppose that this fact is relevant to their moral status. There are several reasons, however, not to write off the importance of the one-shot game in ethical theory. First, not all morally significant actions are repeated. It is difficult to imagine genocide, for example, modeled as a move in an iterated game. Furthermore, in standard treatments of repeated games, the appropriate strategy depends on the likelihood that the game will continue from one round to the next. When that likelihood approaches zero, the game is identified with the one shot game. Morality does not seem like that. The rule that ought to guide my action does not change because this is the last time I am to be faced with a particular kind of choice. It seems at least as virtuous to abide by the usual moral practices of helpfulness, generosity, truthfulness, etc., on one's deathbed as in the prime of health. Indeed, the kind of "cooperation" in one round of a repeated game that is explained by the rewards and punishments that the players might expect in later rounds is exactly the kind of cooperation that does not require morality at all.

Even those morally significant act types that we do perform repeatedly are often performed only once with a particular partner and with no knowledge of what moves that partner has made in previous play with others. Consider, for example, Ann's refraining from stealing the coins in the cup of a blind beggar she passes on vacation in San Francisco, or Baden's gallant efforts to help an old lady across the street. It is unlikely that Ann or Baden knows anything about the prior "play" of their partners, or that the pairs will interact again with similar or switched roles. This suggests that when we choose to employ the iterated game, it would be appropriate to consider a version in which the history of previous play is unavailable to the players. Strategies like *Tit for Tat* that are conditional upon play in previous rounds then make no sense. Since the relative payoffs for unconditional strategies are the same no matter how many iterations of a game are played, we may as well take each stage to comprise just one (possibly mixed) play of the game by every pair of members of the population. The games can still be considered evolutionary. After each stage, the number of players playing successful strategies increases relative to the number playing unsuccessful strategies. The stages themselves, however, are viewed as sets of one-shot games being played simultaneously rather than repeated *sequences* of games. To distinguish the framework suggested here from the one that has become more commonly adopted, we call the former *parallel* and the latter *serial*. Parallel evolutionary games were, in fact, employed by Maynard Smith and others in early biological applications of game theory and emphasized in the writings of Bryan Skyrms and Robert Sugden. But it in more recent literature they seem to be increasingly neglected in favor of serial games. One-shot games, parallel evolutionary games and serial evolutionary games may all clarify aspects of moral philosophy, and I will have things to say about all of them. Since they have different properties, it will be important to specify which kind of game is being considered.

2.2. The dynamics of evolution.

Many properties of evolutionary games are sensitive to the choice of *evolutionary dynamics*, i.e., of the rules that determine how the population of players is altered at each stage.

For biological purposes it makes sense to think of the players at each stage as comprising a *generation* and the payoffs as measures of *fitness*, indicating the expected number of like-playing offspring a player will have in the next generation[3]. If we assume that the total population remains constant, this gives rise to the *replicator dynamics* described by the following equation:

R) $p_i^* = p_i(V_i/V)$

Here p_i*, the proportion of the population playing strategy i at a given stage, is the product of the proportion p_i playing i in the previous stage and the ratio V_i/V of the payoff of strategy i to the average.

When the concern is cultural transmission rather than biological, it makes sense to think of a fixed population of players who discard and adopt various strategies as the game progresses, rather than a changing population whose members, on death, leave varying numbers of offspring. It is not clear that the replicator dynamics is appropriate in these contexts. Those who bother to discuss their use of replicator dynamics for models of cultural transmission sometimes point to results establishing that replicator can arise under various other scenarios of player choice. In one scenario (described, for example, in Gintis), players compare their own payoffs in the current round with the payoffs of others in that round. Each player i switches to the strategy of player j with probability zero if $V_i < V_j$ and with probability proportional to V_j-V_i otherwise. In another scenario (described in [Bendor and Swistak, 1997]), player i's probability of switching to the strategy of player j depends on two attributes of j: j's payoff relative to the average, V_j/V, and the proportion N_j /N of players employing j's strategy. The first parameter is viewed as an indicator of the present success of j's strategy and the second parameter as an indicator of its past success. We suppose, further, that i's probability of switching to j's strategy rises exactly as much when the first parameter is raised by a fixed amount as when the second parameter is raised by that amount, and that the rate at which i's probability of switching to j's strategy grows with the proportion of players using j is itself independent of that proportion. Under both of these scenarios, it turns out that the replicator dynamics correctly describes how the population evolves, and this may be thought to justify the common practice of borrowing this biological model for descriptions of cultural evolution.

Several authors have suggested that a more appropriate dynamics for cultural transmission is *winner imitation*. After each generation, every player whose payoff is less than anybody else's switches (with probability) to a strategy whose payoff is maximal. In the version of winner imitation described in the equation below, it is assumed that if there is more than one strategy with maximal payoff, the players who switch choose randomly among the winning strategies and therefore divide themselves evenly among them.

$$\text{WI)} \quad \overset{\bullet}{p_i} = \begin{cases} p_i + \frac{\alpha}{K}\sum\{p_j : V_j < V_M\} & if V_i = V_M \; and \; K = \left|\{j : V_j = V_M\}\right| \\ \bar{\alpha} p_i & if V_i < V_M \end{cases}$$

Let us temporarily set aside the question of whether any of these stories about evolutionary dynamics are appropriate for serially repeated games. In the parallel framework, especially if some mixed strategies are permitted, they seem implausible. If players don't know the history of others' past play, it is not likely that they know the strategies that have produced the current behavior, and so it is unlikely that they will be able to imitate more successful strategies. A more plausible story is this. Players are more likely to stick to a strategy if it has been successful and they are more likely to switch to another strategy (*any* other strategy) if it has not. So the probability of changing to a new randomly chosen strategy in each round should be a decreasing function of one's relative payoff under the old strategy in the preceding round. We might label such a dynamic "failure-induced groping," or FIG. A FIG dynamics is still an evolutionary dynamics in a broad sense: the numbers of the successful strategies tend to increase relative to the unsuccessful ones. One reasonable FIG dynamic, for example, is described by equation below (where N is the total number of strategies).

$$\text{FIGp)} \quad p_i^* = p_i(1-\alpha\frac{V_M - V_i}{V_M - V_m}) + \frac{1}{N-1}\sum\{p_j\alpha\frac{V_M - V_j}{V_M - V_m} : j \neq i\}$$

Here the probability of abandoning strategy i in favor of a randomly chosen strategy is proportional to the size of i's payoff relative to the largest and smallest payoffs obtained by any players. Specifically, the probability of abandoning i is proportional to $(V_M\text{-}V_i)/(V_M\text{-}V_m)$ where V_M and V_m are the largest and smallest payoffs obtained. If the constant of proportionality α is one, a player who gets the maximum payoff will maintain the current strategy with certainty and a player who gets the minimal payoff will abandon the current strategy with certainty. If α is less than one, the players will be less likely to abandon inferior strategies. It is assumed that players who abandon their strategies distribute themselves evenly over the remaining strategies. Another plausible FIG dynamic, described by the equation below, assumes that each player wants to maximize his expected payoff.

$$\text{FIGd)}\ p_i^{\bullet} = \begin{cases} p_i + \dfrac{1}{N-1}\sum\ \{p_j : V_j < V - \beta\} & \mathit{if}\ V_i \geq V_j - \beta\ \mathit{and} \\ \dfrac{1}{N-1}\sum\ \{p_j : V_j < V - \beta & \mathit{if}\ V_i < V - \beta \end{cases}$$

To this end he retains his strategy with certainty if its payoff exceeds the average payoff in the population and abandons it with certainty otherwise. Since there may be some cost to switching, the equation requires that i's payoff exceed the average by more than some threshold constant β. The first FIG dynamic is *proportional*, the odds of switching increase continuously as payoff declines. The second is *discrete*, the odds of switching jump from 0 to 1 when conditions warrant. Both FIG dynamics require that players know something about the payoffs to others in their generation, but neither requires that they know the strategies others used to obtain those payoffs.

If *all* mixes of permissible strategies are permitted, a plausible dynamic might involve *stochastic learning*. Each player observes the strategy that is realized when she implements her mixed strategy and the payoff that results. She updates her mix noting the component of her mix that actually gets realized in play and by adding to the weight of the component realized in proportion to the relative payoff received.[4] Alternatively one might consider a *stochastic replicator* dynamics, whereby each player changes the weights on his mix in proportion to the relative payoffs received by the strategies realized by *others* in the previous round.

Each of the dynamics described can and should be modified to include the possibility of *mutation* or *error*. In each generation a few players may be assumed to change to (or give birth to) a strategy chosen at random from the strategies represented in the population. This feature may significantly change the course of evolution. Without it, for example, none of the above dynamics except the stochastic ones would allow the population to change once every strategy received the same payoff. Although some writers have done so, it is not reasonable to suppose that the rate of such mutation or error is zero. The idea of error or mutation suggests that an unexpected player replaces a particular member of the old population. Sometimes, however, mutants are described as "invaders," suggesting that they are just added to the population, and the dynamic, "makes room" for them by a tiny proportional adjustment in all the other strategies. This distinction should not matter unless the population size is small.

The FIG dynamics and the stochastic learning dynamics employing relative payoffs place fewer informational requirements on players than the replicator and winner imitation dynamics that have been more widely used. For purposes of modeling moral evolution, it may turn out that even these requirements are too strict. Some even less demanding dynamics will be discussed in subsequent sections. It may well turn out that, for games of interest to us, a variety of these dynamics are equivalent in the sense that they take identical initial populations to identical final states. Until that has been demonstrated, however, we should study the dynamics most plausible for the application we have in mind. And even after it has been demonstrated, the more faithful dynamics might be retained to understand the path from initial to final state or the speed with which that path is traversed.

2.3. Solution concepts

One of my exhortations is to think more carefully about conditions that a set of strategies must meet to be considered a solution to a game. Formulating such conditions requires reflection on the phenomena that the game is supposed to model and the insight the modeling is supposed to provide. In the evolutionary context, we generally think of these games as modeling purely descriptive phenomena: what people actually do or believe rather than what they should do or believe. The observation that the strategies we employ constitute a "solution" to such a game should then serve as an *explanation* for our employing them. Players don't really "solve" these games themselves, but instead evolution leads them inevitably to a “stable equilibrium.” The trouble is that common notions of evolutionary stability are often either too strong or too weak. If they are too strong, they will not be met by any population of strategies--including the one describing the state whose presence we wish to explain. If they are too weak, they will be met by too many populations of strategies. The fact that our state happens to correspond to one of these will not have much explanatory force. Some details of this awkward situation are summarized below.

Originally, evolutionary stability was taken to be a property of strategies. A strategy is supposed to be evolutionarily stable if a population using that strategy will not be dislodged by the course of evolution. A number of authors formulated conditions on game payoffs that were intended to characterize this notion of a stable strategy. Early work on evolutionary games was marred by a failure to realize that these characterizations are not equivalent. For example, Axelrod and Hamilton claims to show that Tit for Tat was an evolutionarily stable strategy in the sequentially repeated prisoner’s dilemma, while Selten and (Boyd and Lorberbaum) offer proofs that no such strategies exist, all apparently unaware that they are employing three different concepts of evolutionary stability.

(Bendor and Swistak, 1998) does a good job of sorting out the confusion. Listed below are conditions for evolutionary stability of strategy i employed by Axelrod, Maynard Smith, Boyd and Lorberbaum and Bendor Swistak. (j is a strategy and V(i,j) is the payoff to a player employing i when meeting a player employing j.)

A) j[V(i,i) V(j,i)],

MS) j[V(i,i)>V(j,i) or (V(i,i)=V(j,i) & V(i,j)>V(j,j))],

BL) j[V(i,i)>V(j,i) or (V(i,i)=V(j,i) & kV(i,k) V(j,k))],

BS) j[V(i,i)>V(j,i) or (V(i,i)=V(j,i) & V(i,j) V(j,j))].

MS and BL each correspond to a notion of *universal* stability: strategies meeting these conditions survive invasion under *any* evolutionary dynamic. BS corresponds to a notion of *restricted* stability: strategies satisfying it survive under the replicator dynamic, but perhaps not under other evolutionary dynamics. MS is *strong*: strategies meeting it drive to extinction any mutations that appear among them; whereas BL is *weak:* such mutations could survive among the natives, even though they do not grow. MS is *narrow*: strategies meeting it are invulnerable only to homogeneous invasions of mutants, whereas BL is *broad*: strategies resist mixed invasions as well. (The distinction between broad and narrow stability is not relevant for restricted stability: under the replicator dynamics a strategy resists heterogeneous invasions if and only if it resists homogeneous invasions. Condition A merely states that i forms a nash equilibrium with itself. If i satisfies A then a population of i-players will not be dislodged under an evolutionary dynamic with a zero mutation rate. It may be overthrown, however, by a single invader. A population composed entirely of a strategy i that meets A, but not the other stability conditions is sometimes said to be an unstable equilibrium.

There is no good reason, in discussing concepts of stable equilibria, to restrict attention to populations playing a single strategy. Solution concepts in non-evolutionary games, after all, do not require that players play similarly. We may as well take any strategy set (i.e., any partition of the population into subpopulations by strategy) to be evolutionarily stable if it is similarly impregnable. The definitions above require some modification in the broader framework. No heterogeneous population can really admit invaders because the strategy set representing the original population is changed, (and thus “driven extinct”) by the presence of a single invader. It is more appropriate in the general context to say that a population is strongly stable if the dynamics restores the strategy set to its original condition after a small invasion, and weakly stable if the dynamics does not carry the strategy set farther from its original condition after a small invasion.

One might expect that all the stability concepts from (Bendor and Swistak, 1998) discussed above would easily generalize to the broader framework. Replace each occurrence of strategy i in the above conditions with a strategy set *profile* $=(s_1,\ldots,{}_{Sn})$ where $s_1,\ldots,{}_{Sn}$ are the proportions of the population playing strategies 1,…,n, and take V(,x) to be the *average* payoff to a member of when each member of plays x (if x is a strategy) or when it plays each member of the population represented by x (if x is a profile.) This does not quite work in finite populations, however. Consider, for example, a two-person two-move game whose only nash equilibrium is a fifty-fifty mix of the two moves. One example is *penny mismatch*, where each player chooses "heads" or "tails," with the object of choosing a different face than his opponent. An evolutionary version of this game played with n players (who cannot themselves mix strategies) has a profile meeting MS if n is even (viz., (.5,.5)), but, if n is odd, it doesn't even have a profile meeting A. If we make some reasonable stipulations,[5] then the closest attainable approximations to (.5,.5), namely ((n-1)/2n, ((n+1)/2n) are, in fact stable. More generally, we can take the simple generalized stability conditions to apply to "games" in which there are a continuum of players so that all profiles $(p_1,\ldots,p_n)$ 0 p_i 1 are realizable. To find the equilibria in the n player game, we first find those of the game with a continuum of players. Then we look at what happens to the closest approximations to these in the n-player game. If the dynamics carries one of these profiles to its neighboring equilibrium, it is itself an equilibrium in the n-player game. Otherwise, it is not. If the equilibrium it is carried to is stable, so is the approximation, otherwise it is not. Examples of these phenomena are given in the appendix. In what follows, therefore, we generally take the stability conditions to apply to strategy profiles and assume a continuum of players, with the understanding that information about the n-player case can be extracted.

Once the stability conditions are set out it is easy to map the logical relations among them. MS and BL are independent and each implies BS, which in turn implies A. For sequential evolutionary games, MS and BL are too strong: they cannot be met. BS and A are generally too weak. For example, in an evolutionary version of the iterated prisoner's dilemma they are met by populations supporting zero percent cooperation, one hundred percent cooperation or any figure in between.

One source of the "excessive strength" difficulty is that sequentially repeated games allow so many different strategies. For example, the reason that MS cannot be satisfied by any single strategy is that for any strategy i, it is possible to construct a strategy j different from i that mimics the way i plays against i and j. Clearly in this case we can't satisfy either of the inequalities V(i,i)>V(j,i) or V(i,j)>V(j,j). So the difficulty may not arise when the permissible strategies are limited. This is exactly what happens in the case of the parallel versions of the evolutionary game. These can be regarded as obtained from the sequential game by allowing only the simplest strategies–the unconditional ones.

If there is any strategy i such that (i,i) is a strict nash equilibrium in the underlying one-shot game, i.e., any i such that V(i,i)>V(j,i) for all j i, then i satisfies every stability condition listed above. Conversely, if i meets any condition on the list then (i,i) is a weak nash equilibrium in the underlying game. Among the strategies that form weak nash equilibria with themselves, some meet MS but not BL some meet BL but not MS, and some meet neither. (Examples are provided in the appendix.) Since many parallel games have strategies that form nash equilibria with themselves, the difficulty of excessive strength may not be a great concern in this framework.

To meet the difficulty of excessive weakness, at least two ways of discriminating among the stable profiles have been invoked. Suppose S is a stable profile. The "basin of attraction" of S under a particular rule of evolution is the set all profiles that evolution carries to S. The "robustness" or "degree of stability" of S is the size of the smallest invasion that can overturn S. The strategies with the largest basin of attraction need not be the ones with maximal degree of stability. For example, consider the profiles for a game where each player can play either *left* or *right*. It is conceivable that the evolutionary dynamics carries all profiles (0,1) through (.6,.4) to an equilibrium at (.5,.5) and all other profiles to an equilibrium at (.8,.8). Then the basin of attraction of (.5,.5) would be larger than that of (.6,.4). But the former equilibrium could be upset by an invasion of *right* players comprising ten percent of the population, whereas the latter would require an invasion of 20%.

Which is a better explanation for our being in a particular state S, the observation that S has the largest basin of attraction or the observation that S has the largest degree of stability? The best explanation for our being in S would seem to be the observation that S is where we are *expected* to be. The basin of attraction would only be a good explanation if we thought that every initial distribution of strategies was equally probable and that evolution stopped as soon as an equilibrium was reached. In fact, however, neither of these conditions is plausible. First, if we assume that each player starts with a randomly selected strategy, then the equal initial distributions are much more likely than the unequal ones. Second, as was noted previously, it is not reasonable to suppose that offspring never differ from their parents in the biological model or that there are never “accidental” strategy shifts in the social modal. In either model, mutations continually threaten the status quo. Under these conditions’ one would expect the length of time that a population spends at any stable point to correlate with the number of mutants required to upset it. So our being in a particular state is best explained by the degree of stability of that state.

2.4. Interpersonal Comparisons and Asymmetry

The evolutionary dynamics and the conditions on payoffs characterizing equilibria discussed above all seem to require that payoffs be interpersonally comparable. According to the dynamics described in section two above, the growth or decline of a particular strategy under non-equilibrium conditions depends on comparisons between the payoffs to players who have adopted that strategy and those who have adopted others.[6] According to the solution concepts described in section three, the achievement of equilibrium depends on comparisons between the payoffs of natives and invaders. When our concern is biology, and payoffs are just measures of reproductive success, this is quite appropriate. When our concern is culture, however, we should be a little more wary. It is common to regard the payoffs of these games as utilities, and the questions of whether interpersonal utility comparisons are meaningful and measurable are notoriously vexed.

Those writing about evolutionary games who bother to discuss the issue typically deny that their framework requires interpersonal utility comparisons.[7] One understanding that might make this possible is that all the references to payoffs of others in a given environment are understood *counterfactually* as the payoffs that one would oneself get in that environment, if one adopted the other’s strategy. Thus, for example, in the replicator dynamics Player j is more likely to adopt i’s strategy when his payoff under his current strategy is less than what he himself would have gotten had he used i’s strategy. Equilibrium is achieved when each player “sees” within the population no strategy that will serve him better.

What makes this counterfactual interpretation possible is a special kind of symmetry that has, until now, been built into our framework. Payoffs to players depend only on the strategies they and their opponents adopt. This assumption is what makes our payoff notation coherent. V(i,j) is the payoff to *any* player adopting strategy i against *any* player adopting strategy j. It follows, of course, that exactly the same payoffs are available to all players and one player can approximate another's payoff just by adopting her strategy.

Interpreting payoffs counterfactually in familiar conditions for evolutionary dynamics and stability raises certain questions. First, while player i can approximate j's payoff in a particular environment by adopting j's strategy he may not be able to duplicate it, because i's playing his current strategy is a part of the environment in which j was acting. Of course if there are many players, the effect of play against any one opponent on the payoff for the round will be negligible. Nevertheless, if we really take seriously the idea that the payoffs are to be interpreted counterfactually, then, instead of comparing V_i with V_j in the current round, i should compare V_i with the payoff that a player would get by employing j's strategy in an environment where one fewer player uses i's strategy and one more player uses j's. Second, the interpretation seems to render the replicator dynamics less plausible. As long as we take each player to be influenced by others' payoffs, we can imagine that the degree of j's influence on i is proportional to the size of j's payoffs relative to i's. But if we take i to have full knowledge of what his payoff would have been under every strategy currently represented in the population, it is difficult to see why i wouldn't simply choose the highest paying option. Of course i may realize that others are reasoning similarly and so the environment will change, but that would seem to provide no justification for the particular weighting of lower-paying alternatives called for by replicator. Similarly, if i knows what his payoff would have been under every strategy currently in the population, it is difficult to maintain that he doesn't also know what his payoff would have been under strategies that are not any longer in the population. So why shouldn't i choose a strategy that maximizes his payoff whether it is represented in the population or not? One might hold that players do not know what payoffs they would get under various strategies, but rather infer this by observing the payoffs of others who use them. But this would require players to know that the others' payoffs are the same as they themselves would reap by adopting their strategies.

In other words it would require that the players themselves can make interpersonal comparisons.

I conclude from all this that the plausibility of standard accounts of evolutionary dynamics and equilibrium requires that it makes sense for one player to compare his payoffs with another. If we are really concerned about the use of interpersonal comparisons it would be reasonable to adopt an evolutionary dynamic under which each player's probability of switching strategies in a given round depends on his own relative payoffs in this round and previous rounds. Each player strives to do as well as he *can* do. He does so by changing frequently when he does worse than he *has* done and occasionally (i.e., at the "mutation" rate) even when he does as well as he has. The details could be worked out in a variety of ways. For example, we could devise a scheme whereby players weigh payoffs in recent rounds more heavily than those in earlier rounds when deciding whether current payoffs are low enough to warrant switching strategies.

It is possible to establish a few properties of equilibria under this kind of "individualistic" dynamic without knowing how it is specified. First, such equilibria really do not require interpersonal comparisons: a population in equilibrium will remain in equilibrium if the payoffs some of its members undergo monotone transformations. Stable populations need not get equal payoffs. They will form a general nash equilibrium, however, in the sense that each player's strategy is a best reply to the configuration of others employed in the population. For otherwise the player whose strategy was not a best reply would keep switching (by "self-comparison" and "mutation") until he did reach a best reply. Not every nash equilibrium is stable, however. A deviation by one player from a nash equilibrium may lead to a deviation by another that leaves everybody better off than they were originally. Any solution in which each player gets her maximum payoff is, of course, stable. Indeed, if the dynamic is specified appropriately, any unique nash equilibrium will be stable. When players have reached such a state, some of them may initially "remember" higher payoffs. Attempts to recapture this payoff, however, will fail and eventually the memory will fade. If there is more than one nash equilibrium, the relative stability of each might be measured, as in the case of the interpersonal dynamics, by the number of strategy changes needed before the dynamics carries the population away from that state. In the appendix we note that, for the game with a continuum of players the condition that a profile forms a nash equilibrium is the same as the condition that it meets A. So the individualistic dynamics may not be so different as it appears.

For the most part, I will assume in what follows that interpersonal comparisons of payoffs are legitimate and adopt the conceptions of evolutionary dynamics and stability that were discussed above. In the final class of examples, however, I will consider games that violate the special symmetry assumption. Players will be divided into types (privileged and deprived) and the payoff to i for playing j will depend on i's type as well as the strategies of i and j. For such games the idea of adopting strategies that others find rewarding seems particularly inappropriate, and so I will consider solutions that would be reached by the alternative kind of dynamics.

3. The Examples

3.1. Suboptimal solutions

Let us now turn to the challenging examples. The existence of the first class has been noted from the very beginnings of game theory. These are games with one outcome that constitutes a solution according to the solution concept appropriate for the intended application and another outcome that all the players prefer. The most notorious example is the familiar two-person one-shot prisoner's dilemma. In this case there is only one nash equilibrium, mutual defection, and yet both players prefer another outcome, mutual cooperation, that is not an equilibrium at all. There are other games in which the unanimously preferred outcome *is* an equilibrium. One example is a version of the Stag Hunt, a game that has recently been commended to our attention as an important representation of the social contract. (See [Skyrms, 2001].) Consider the payoff matrix below.

Stag Hunt

	Stag	Hare
Stag	3,3	0,2
Hare	2,0	2,2

Row and Column each choose between hunting stag and hunting hare. A successful stag hunt requires participation from both and provides a payoff that both prefer to a successful hare hunt. Success in the hare hunt, however, does not depend on the participation of the other. If both choose to hunt hare, it would be futile for either to switch prey. If both choose to hunt stag, then either would do somewhat worse by switching prey. Thus, both hare hunting and stag hunting are strict nash equilibria in the one-shot version of this game. Stag hunting, of course, is unanimously preferred to hare hunting, but we would expect it to be reached only if each player is reasonably confident[8] that the other will choose to hunt stag.

Since (stag, stag) and (hare, hare) are both strict nash equilibria in the one-shot stag hunt, everybody's hunting stag and everybody's hunting hare are both equilibria in the strongest sense in the parallel evolutionary version of the game. A population of stag hunters will eradicate small invasions of hare-hunters or mixers under any rule of evolution. A population of hare hunters will similarly quash invasions by stag hunters. Closer examination, in fact, reveals that the "inferior" equilibrium of hare hunting is more stable than the stag hunting equilibrium. A population of hare hunters will overcome stag hunting invaders until the invaders reach two thirds of the population[9], whereas a population of stag hunters will resist invasion by hare hunters only while the hare hunters comprise less than one third of the population. So, if solutions to the parallel game are maximally stable equilibria, the parallel evolutionary stag hunt, like the one-shot prisoner's dilemma, is game whose solution is suboptimal.

The problem of suboptimal solutions to parallel repeated games can arise even when the game does not have exactly the stag hunt structure. Suppose, for example, that the hare hunter's chance of success also increases with cooperation, so that the payoff for hunting hare alone is 1.5 rather than 2. Then we have a "coordination game": both players always do strictly better when they perform the same action than when they perform different actions. In this case the hare hunting equilibrium is stable under invasions comprising up to about 57% of the population whereas the stag hunting equilibrium is stable only under invasions comprising up to about 43% of the population. Or suppose that the hare hunters are *competing* for limited resources, so that the payoff for hunting hare alone is 2.5 rather than 2. Then we have a slightly different game structure: both players always benefit by choosing the same action as their opponent rather than a different action, but they also always benefit when their opponent chooses stag over hare. In this case the hare hunting equilibrium is stable under invasions up to 80% compared to only 20% for the stag hunting equilibrium. All three examples are instances of what has been labeled (in Sen) “assurance” games because, unlike the PD, the “cooperative” outcome can be reached if each player has assurance that his opponent will cooperate. Viewed as parallel evolutionary games, all three are examples in which another outcome is unanimously preferred to the most stable equilibrium.

All three games considered above (as well as the PD), satisfy conditions defining what Bendor and Swistak call "games of cooperation," where hunting stag is considered the "cooperative" move. (Bendor and Swistak, 1997) shows that, in two-player repeated serial games of cooperation under the replicator dynamics, no strategy can resist invasions comprising more than 50% of the population. If the discount rate is sufficiently low, there are a number of strategies that approach this maximum possible degree of stability. One is the celebrated Tit for Tat, which in this context says "hunt the prey that your opponent hunted the last time you played him." More importantly for our purposes, they show that all maximally stable strategies are "almost nice," which means that they "almost always" achieve the cooperative payoff. So for serially repeated games that evolve by the replicator dynamics I have given no examples of suboptimal solutions. I am not sure whether there are any such examples or not. In view of the initial remarks, however, this should not provide much comfort. For many of the kinds of games that are most relevant to ethics, it is possible to reach "solutions" in which everybody is worse off than they might otherwise be.

What exactly is the "challenge" that suboptimal solutions pose to game theoretic treatments of ethics? The answer depends on what we take ourselves to be investigating. Suppose first that we take our subject matter to be genuinely normative. Solutions mark patterns of behavior that we really ought to follow. In this case a suboptimal solution marks a situation in which everybody does the right thing and yet everybody would be better off if they did something else. There are no limits in principle to the size of the gaps that might arise between the solutions. We might find ourselves in a situation in which each of us suffers enormous harm because we all do A and each of us would gain enormous benefit if we all did B. Even if we don't accept the strong connection between duty and welfare espoused by utilitarians, it seems implausible to suppose that it is right to do A in this case. For that would imply that morality is an institution that we would *all* be better off without. More specifically it would imply that we would all do better to replace morality with a variation that asks us to depart from morality exactly in those conditions when we face suboptimal solutions.

A natural way to meet this challenge is to distinguish between normative concepts for groups and individuals, and between conditional and unconditional normative concepts. *We* are wrong to do A, but given that the others do A, I am right to do so as well. These distinctions are undoubtedly significant and useful in this context. It is important, however, that making them does not allow us to evade the serious moral questions that are raised by the class of examples. What should I do when faced with a choice between A and B? To what degree should my choice be influenced by my expectations about what others will do or by my convictions about what they should do? How much credit or blame do I deserve when *we* do something right or wrong?

Suppose next that we emphasize descriptive interpretations of solutions. Skyrms, for example, sees his work as part of a tradition, including Hume and Rousseau, that seeks to investigate the evolution of existing patterns of behavior or, as he puts it, the evolution of "an existing implicit social contract" rather than the sort of *ideal* social contract contemplated by Rawls, Harsanyi, or Hobbes. Likewise, Sugden sees himself as showing "that certain kinds of conventions tend to evolve spontaneously in human society" (p172). Sugden takes the further important step of identifying a subclass of these conventions that "acquire moral force," i.e., that *come to be regarded* as moral conventions.[10] He explicitly disavows a purely normative interpretation of these conventions.

> "It is no part of my argument that the morality that evolves in human society is the morality we ought to follow. I am not trying to present a moral argument; I am trying to explain how we come to have some of the moral beliefs we do." (p 175)

I believe that the game theoretical investigations do have normative import. It seems implausible that a theory that explained exactly how we came to hold the moral beliefs that we do would have no implications about whether we should act on them. This thesis, however, is independent of Skyrms' and Sugden's reminders that there is another interesting and important subject of study. We might distinguish between "descriptive ethics" and "prescriptive ethics," just as P.F. Strawson, a generation ago, distinguished between "descriptive metaphysics" and "prescriptive metaphysics." So let's assume, for now, that our concern is with the former rather than the latter.

The identification of suboptimal solutions with morally correct patterns of behavior seems curious even granting a descriptive interpretation of "morally correct." It is easy to imagine that we might all believe we ought to do A, not realizing that we'd all be better off if we did B. It is more difficult to imagine that we would retain this belief on becoming aware of B's advantages. One of our general moral beliefs is that it is possible that patterns of behavior that have become widespread in our society are *wrong*. Surely a fascinating part of descriptive ethics to which game theory might be expected to contribute is the phenomenon of moral *change*. Behavior that was once considered lewd is now viewed as acceptable or entertaining. Jokes that were once considered funny are now considered insensitive or racist. Many like to think that there has been moral progress. It seems reasonable to suppose that at least one form of moral progress consists in moving from one equilibrium to a unanimously preferred one. Suppose, in our example above, doing A is a generally accepted norm in some society. It is very easy to imagine moral reformers successfully urging its members to change: "We should not really be behaving this way. We should be trying to establish B as the norm." It is much more difficult to imagine moral reformers successfully urging its members to move from B to A. This property of *irreversibility* of change seems to be a hallmark of what we consider moral progress.

One answer to our questions about suboptimal solutions is suggested in the writings of Kenneth Binmore. For Binmore the application of game theory to ethics is a two-stage affair. Solutions of the sort we have been discussing are equilibria in the "game of life." They indicate which patterns of behavior it is possible to sustain. The "game of morals" is the game by which we select one among the various equilibria. Binmore's idea works very nicely for the coordination games like our stag hunt examples. Whatever the correct principles of equilibrium selection turn out to be, it is reasonable to suppose they will choose an equilibrium that benefits everybody over one that benefits nobody. It is far less plausible, however, for the cases like the one-shot prisoner's dilemma in which the preferred outcome is not an equilibrium. For Binmore, defection is the only moral choice in a one-shot prisoner's dilemma. Yet for philosophers like Gauthier and Kurt Baier, cooperation in a one-shot prisoner's dilemma is paradigmatic moral behavior: the prisoner's dilemma serves to explain how moral rules can be "advantageous for everyone" while requiring "that some persons perform disadvantageous acts." Binmore defends his insistence that equilibria are the only possible candidates for moral behavior as a consequence of the principle that "ought implies can." That argument, however, overlooks the significant fact that inculcating moral beliefs changes the payoffs, so that behavior that is not in equilibrium absent moral beliefs, may well be sustainable in the presence of those beliefs.

The solution to the problem of suboptimal equilibria (or at least to those varieties of the problem illustrated by the prisoner's dilemma or stag hunt) favored by Skyrms and others appeals to *correlation* of interactions. Players do not play against all opponents in a round of a parallel evolutionary game or with a random sample of such opponents. A general theory of evolution under correlated interactions presents some difficulties (Skyrms, 1994), but particular cases are straightforward. On a *spatial* model ([Nowak and May], [Nowak, Bonhoeffer and May], [Grim, Mar and St Denis], [Skyrms and Alexander]), players are arranged in a fixed pattern. Each player interacts only with its *neighbors* and considers only its neighbors payoffs when in deciding whether to switch strategies. On a *social network* model (Skyrms and Pemantle) players are more likely to interact with those with whom they have had previous beneficial interactions. Spatial models make it more likely that a player will interact with someone using the same strategy as herself. With appropriate implementation, social network models may have the same effect in prisoner's dilemma and stag hunt games. Since (cooperate, cooperate) and (stag, stag) are the better of the two outcomes in which both players use the same strategy, a sufficiently strong correlation makes it possible for the players to escape the suboptimal equilibrium.

The spatial and social network models are certainly suggestive, and they may help to explain certain patterns of human behavior. Those interested in applications of moral philosophy to game theory should be somewhat wary of appeals to correlation, however. The spatial models that are so helpful in allowing players to achieve desirable outcomes in games calling for similar play, like the prisoner's dilemma and stag hunt, will force them into the least desirable outcomes in games, like hawk-dove, that call for different play. More generally, appeals to correlation, like the appeals to reciprocity by purveyors of serially repeated games, fail to adequately explain specifically moral features of behavior. There may be cultures in which moral norms apply only to those with whom a person has frequent interactions. In general, however, we would like an accounts that explain our attitudes towards interactions with outsiders as well as insiders. And we'd like accounts that explain why behavior is (or is perceived to be) morally correct as well as why it happens to be practiced. In his discussion of bargaining games, Skyrms remarks cautiously that correlation, among other factors, "is perhaps a beginning of our concept of justice." But, while these factors may explain our behaving in ways that happen to be just, they do not explain our having the *concept* of justice that we do. To explain that we must, as suggested above, say something about the way we employ moral rewards and sanctions to change the payoffs of the problematic games.

3.2. Mixed populations

The second kind of example I wanted to discuss is the phenomenon Skyrms has labeled the "polymorphic trap." Skyrms raised the issue in connection with a parallel evolutionary version of what he calls the cake division game. A cake is to be divided between two players. Each player requests some fraction of the cake. They both get what they request as long as their requests total no more than the whole cake. Otherwise, both get nothing. Skyrms notes that, for all pairs of fractions (r,1-r) totaling to one, there is some division of the population between those who request r and those who request 1-r that is evolutionarily stable. For example, consider a population of which one-third request 60% and two-thirds request 40%. The greedy players get nothing when they meet each other but they get .6 when they meet a modest. So their expected payoff is (1/3) 0+(2/3) .6=.4. The modest get .4 every time. Since their expected payoffs are the same, the relative proportions of greedies and modests will not change. Now suppose, for example that a small group of players who request 50% tried to enter the population. Since they are few in number, they have virtually no chance of meeting each other. They have a one-third chance of encountering a greedy and a two-thirds chance of encountering a modest. Their expected payoff is (1/3) 0+(2/3) .5=(1/3) which is less than both the greedies and the modests so they get driven to extinction. Now I suspect that the reason that Skyrms calls this phenomenon a "trap" is that the equilibrium in which everybody plays “request 50%” is unanimously preferred to all the polymorphic equilibria (r,1-r). Recall that in our example all the players had an expected payoff of .4. If everybody requests 50%, everybody gets .5. So everybody does better in the equilibrium (.5, .5) than they do in the equilibrium (.6,.4).

We have already discussed the problem of suboptimal solutions. I want to focus here on the issue of polymorphism itself, so it would be good to use an example where the polymorphic equilibrium was *not* suboptimal. The idea that evolutionary processes might lead to stable mixed populations has frequently surfaced in both simulations and theoretical accounts of the repeated prisoners dilemma game. One example was discussed fifteen years ago in Boyd and Lorberbaum. Consider the following four strategies:

TFT: cooperate on the first move and thereafter if opponent cooperated in previous meeting; otherwise defect.

STFT: defect on first move, play TFT thereafter.

TF2T: cooperate on first move or if opponent cooperated at least once in last two meetings; otherwise defect.

D: always defect

In a universe of these four strategies, Boyd and Lorberbaum show that there is only one strongly stable equilibrium, which (assuming the payoffs have the usual "Axelrod" values) is a mix of 98.6% TF2T and 2% STFT. To get the idea of why this is so, compare this mix with TFT. In a population that is 100% TFT, TF2T is what is sometimes called a "neutral mutant," i.e., it behaves just like the rest of the population, so it can infiltrate freely. Every once in a while a mutant STFT enters the population. TFT2T does better against this mutant than TFT does, so the proportion of TF2T tends to increase relative to that of TFT. When the ratio of TF2T to TFT is sufficiently high, the mutant STFT's begin to do better than the TFT's and so the TFT's are driven to extinction. Now consider the 98.6% mix of TF2T with STFT. A TFT who tried to infiltrate would do worse than the TF2T's because it would do the same as TF2T against the majority and worse than TF2T against the minority. It would do marginally worse than STFT against both the majority and minority. The role of STFT is curious here. STFT does very badly against itself, so we would not expect it to predominate in a stable population. But in small proportions it becomes a terrific ally of TF2T against TFT. Notice also that the TF2T/STFT mix is not really a "trap." A homogeneous population playing TFT would get the cooperative reward of three units in every round. In the mixed population, STFT exploits TF2T on the first round, so the former gets a higher payoff and TF2T gets a lower payoff. But after the first round, both strategies cooperate forever. If the discount rate is sufficiently low, their total payoffs approach those of the homogeneous population.

The idea of diversity in a population as a defense against potential invaders is familiar in biology, but it seems very odd as an ethical principle. What would we say about a population in which 98.6% follow one moral rule and 1.4% follow another rule? Could this model a kind of ethical relativism? It is not likely to illustrate a form of *appraiser* relativism[11], if any such view could be made coherent. One would not expect the minority population to judge the majority according to the minority's rules: things would become a lot worse for everybody if members of the majority began converting to their suspicious rule. Perhaps, then, it models *agent* relativism. The entire population understands that the majority should follow their rule (TF2T) and the minority should follow theirs (STFT). Again, this seems unlikely. If agent relativism has any plausibility, it is when different standards apply *within* groups. Eskimos abandon *their* aging parents, while Japanese sacrifice for *theirs*. But the standards here apply to interactions that occur uniformly. My odds of interacting with you to play the repeated PD are the same whether you follow my strategy or the other strategy. Indeed in the repeated PD model there are *no* differences among the players except for the strategy that they employ. So if the strategies were moral rules they would violate a very plausible principle of universalizability. People who are qualitatively identical should be subject to the same rules. The rules that a person follows should not depend on her mere identity.

The Boyd and Lorberbaum example shows that if the strategies available to a population are limited, a dominant stable equilibrium (indeed the *only* stable equilibrium) might be a polymorphic one, in which different players use different strategies. A simple reply to this unwelcome observation is possible. Suppose a dominant polymorphic equilibrium has proportion p_1 playing S_1, p_2 playing S_2, …, p_n playing S_n. Now consider a new kind of player who adopts a *mixed* strategy of playing S_1 with probability p_1, S_2 with probability p_2,…,S_n with probability p_n. Among the strategies S_1,…,S_n, MIXER behaves exactly like the polymorphic mix of S_1,…,S_n. Since none of S_1,…,S_n can invade the polymorphic mix in small numbers, none of them can invade the homogeneous population of MIXERs. Thus, when MIXER is added to the possible strategies S_1,…,S_n the homogeneous population of MIXERS is strongly stable. The polymorphic population, on the other hand, is now merely weakly stable, because it does not drive to extinction an invasion of MIXERS. So, by adding an appropriate mixed strategy, we can always replace a polymorphic equilibrium with homomorphic a one. Furthermore the payoffs to the players will be exactly the same whether they adopt the pure strategy or the mixed one.

3.3. Mixed strategies

Our "solution" to the problem of heterogeneous equilibria raises another concern. To avoid heterogeneous equilibria, we embraced mixed strategies. We replaced population mixes with action mixes. The idea of adopting a random strategy, however, does not seem appropriate in moral contexts. To see the problem more clearly, we focus on a much simpler example than Boyd and Lorberbaum's–the classical two person PD that we have already discussed under the heading "suboptimal solutions."

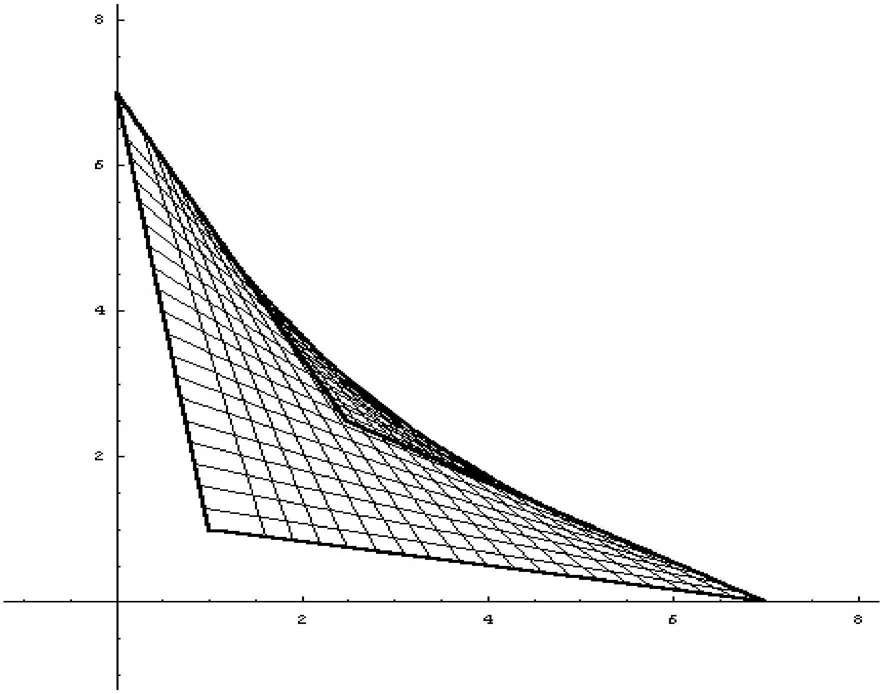
The PD has two important characteristics. First, mutual defection is the only nash equilibrium. In fact defection *dominates* cooperation–it is the best reply not only to the other's equilibrium strategy, but to both of the other's strategies. Second, mutual defection is pareto inferior to mutual cooperation. A fact that I think has been overlooked or underappreciated in discussions of the PD is that–unless a special "purity" condition is met--mutual cooperation is itself not pareto optimal[12]. There is a pair of independent mixed strategies that provides both players with more utility. Consider, for example the PD with the payoff matrix below.

Impure Prisoner's Dilemma

	Cooperate	Defect
Cooperate	2.5,2.5	0,7
Defect	7,0	1,1

The players get two and one half units by cooperating with certainty, which is better than the one unit they get by defecting. If they cooperate with probability 3/4 and defect with probability ¼, however, their expected payoff is $(.75)^2 (2.5)+(.75)(.25)(0)+(.25)(.75)(7)+(.25)^2(1)$ which is approximately 2.78.

The point is made more vivid in the graph in figure 1 below. The payoffs achievable by pure strategies are represented by the four corner points of the concave quadrilateral. If one player adopts a mixed strategy while the other plays a pure strategy, all of the points along the lines connecting these points become possible payoffs. If both players mix all the points in the interior of this quadrilateral are added, as well as the other crosshatched points northeast of quadrilateral boundary. Some of these points are northeast of (R,R) indicating that both players are better off than they are under certain cooperation.



Students of game theory, of course, are well aware that game solutions often call for mixed strategies. The challenge for those who want to apply the theory to moral philosophy is that moral rules do not. No moral philosopher could plausibly suggest that we ought to consult a randomizing device before deciding whether to kill a child or whether to keep a promise[13]. Even in difficult cases like that of Sartre’s patriot, who must choose between helping his invalid mother and joining the French resistance, it seems wildly implausible to think that the decision should be made by a coin flip. I have suggested elsewhere (see Kuhn) that the lesson to be learned from this challenge is that moral philosophers ought to distinguish two subjects–a theoretical one that tells us what we actually ought to be doing, and a practical one that tells us what we should *try* to do. The theoretical subject may be of interest to philosophers, but the practical subject is the one that should interest all of us. This view is a form of what has been labeled indirect morality. How can we best get the effect of, say, 96% compliance with a rule for promise keeping? Surely the best means would not be for each of us to try to keep our own promises 96% of the time, and try to see to it that others do as well. It would be much more efficacious for us each to *try* with a certain degree of effort to keep *all* of our promises, and see to that others do as well.

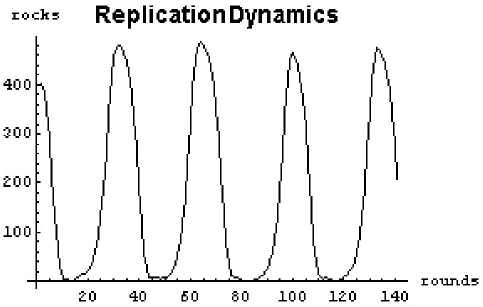
3.4. Cycles

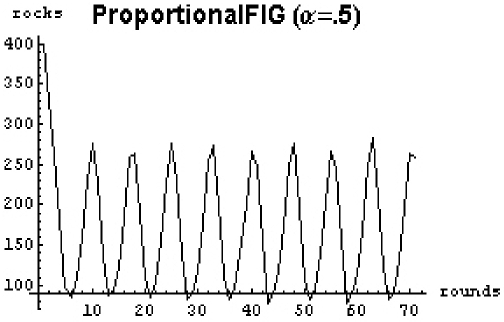
In the previous examples evolution always carries the population to a state with a single dominant strategy or to one with a stable mix of strategies. The next examples are the ones in which evolution leads instead to a repeating cycle of strategies. The existence of this phenomenon is demonstrated by the Rock Paper Scissors game with the payoff matrix below.

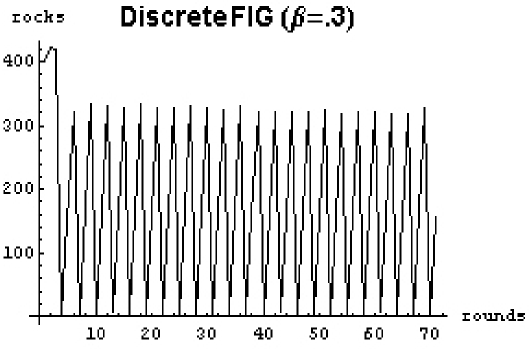
Rock Paper Scissors with Evolutionary Cycles

	Rock	Paper	Scissors
Rock	4,4	1,6	6,1
Paper	6,1	4,4	1,6
Scissors	1,6	6,1	4,4

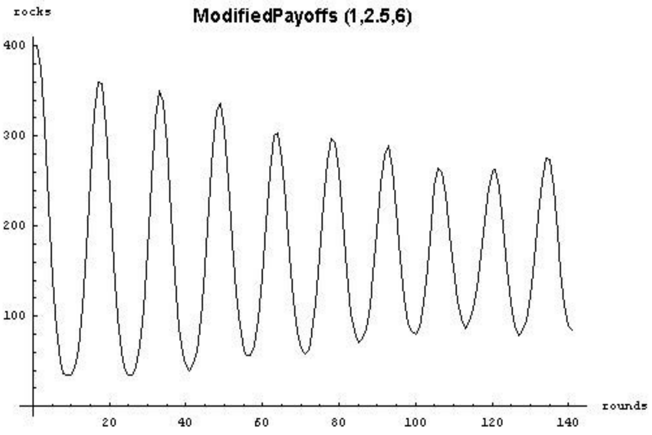
Figure 2 shows the evolution of a population playing a parallel evolutionary version of the game under various rules of evolution. In each case there is a population of 500 players, initially divided with 400 playing rock and 50 each playing paper and scissors. The y axis represents the number of players playing rock. In the first graph evolution proceeds by the replicator dynamics, in the second, it proceeds by proportional failure-induced groping, and in the third, by discrete failure-induced groping. In every case, it is obvious that the only "equilibrium" that is reached is a dynamic one, in which the number of players playing rock grows and shrinks at regular intervals.

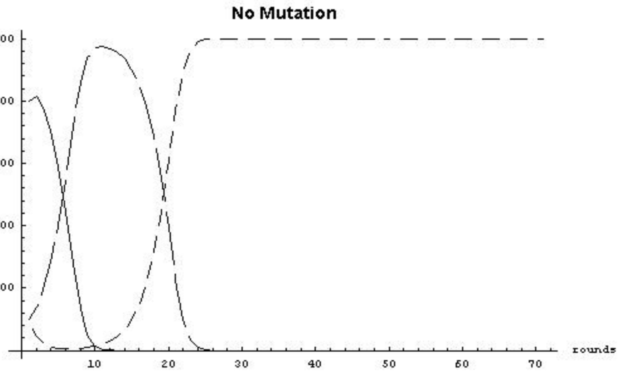

ReplicationDynamics


ProportionalFIG (α=.5)


DiscreteFIG (β=.3)

It should be noted that games like this do not appear to be common. Figure 3 shows what happens when some of the parameters are changed. In the top graph, the payoff for a "tie" between two players is lowered from 4 units to 2.5 units. Under these conditions players do better by taking turns "winning" and "losing" than by choosing the same strategy. As the graph suggests, the oscillations in the population of rocks dampen and we head to a polymorphic equilibrium in which 1/3 of the population plays each strategy. In the bottom graph, evolution proceeds by the replicator dynamics with a zero percent mutation rate. Without mutation, one of the three strategies can be driven to extinction and as soon as that happens the "superior" of the two remaining strategies takes over. In the case pictured, the solid line tracks the rock population, the long dashes track the scissors and the short dashes track the paper. The initial abundance of rocks causes the paper population to grow and the scissors and rock populations to shrink. With few scissors to check them, the papers quickly drive the rocks to extinction, at which point the scissors rebound and drive out the papers.



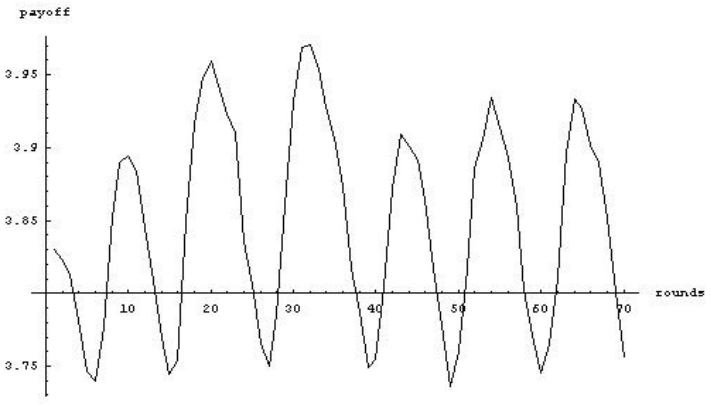


The phenomenon of stable cycles is, of course, quite familiar in the population biology of plant and animal species. Similarly, if our interests are confined to descriptive ethics, then the existence of stable cycles is not problematic. Moral beliefs and attitudes do sometimes appear to change in a cyclical manner. Our attitudes towards drug use and sexual promiscuity, to take two examples, seem to be getting more judgmental in recent years after getting more permissive through the sixties and seventies.

From the perspective of normative ethics, however, the existence of stable cycles is somewhat more puzzling. Those who argue for a return of tighter standards of sexual behavior do not generally maintain that the looser standards were right in their time, while the tighter standards were right before then and again now. They tend rather to call for a return to correct standards, which have been temporarily forsaken. As in the case of suboptimal solutions, it is helpful to distinguish between group and individual concepts of obligation. It is plausible that what I ought to do might depend on what others do, and that what they do varies in regular patterns. When others play scissors I ought to play rock; when others play paper, I ought to play scissors. It is less plausible that what *we* ought to do is to repeatedly cycle though stages during which the proportion of us making a particular choice goes from high to medium to low and back.

One plausible diagnosis of what is going on in cases when moral attitudes change cyclically is that there is some ideal standard that we are unsuccessfully trying to reach. When standards are too strict, we loosen them. In doing so, we overshoot the mark and standards become too loose. Then we tighten them, and the cycle repeats. This story suggests that cycles could be avoided if it were possible to add to the set of possible strategies one that directly expresses the ideal moral standard. In the case at hand, for example, we might suspect that a mixed strategy of 1/3-rock, 1/3-paper, 1/3-scissors, if allowed to arise, would eventually dominate.

This response to evolutionary cycles may sometimes be appropriate, but in this case it is not. Figure 4 plots the average payoff in the cyclical rock, paper, scissors game. The payoff varies cyclically with the makeup of the population but it never goes below 11/3. The mixed strategy, on the other hand, scores exactly 11/3 against every opponent and therefore its average payoff is 11/3. Since the mixed strategy never exceeds the average payoff, it will not successfully invade the cycling population under any of the dynamic rules we have considered. Furthermore, a population initially composed entirely of mixers is vulnerable to an invasion by any single pure strategy. The invaders would do exactly as well against the mixers as the mixers do and they would do better than the mixers against themselves. Once the pure invaders reached a critical mass, of course, they would be vulnerable to attack by the pure strategy that beats them. The cycles would begin.



The general normative lessons to be drawn from the existence of stable cycles in evolutionary games are unclear. There are, however, some similarities between this example and the initial examples of suboptimal solutions. Suppose the evolutionary game is played by a fixed population of players who shift from strategy to strategy. If any of the three strategies were played by all players at all times, each player would achieve a higher payoff than the average she achieves over a cycle in the cyclical outcome. So, just as players defecting in a prisoner’s dilemma have reason to try to move to state of universal cooperation, players locked in a cyclical rock paper scissors game have reason to try to move to a state in which all play the same strategy. As in the prisoner’s dilemma, the state to which they have reason to move is not evolutionarily stable. This is exactly the kind of situation in which we might expect moral attitudes to change the payoffs in such a way that the preferred outcome can become stable. There is a sense in which the cyclical rock-paper-scissors examples are normatively even more problematic than the prisoner's dilemma. In a prisoner's dilemma, we can ensure cooperation if players are moved by appeals to universalizability. When somebody considers defection we ask "What if everybody did that?" In the cyclical rock-paper-scissors game, such appeals are otiose. Suppose we are all playing rock. I can do better by switching to paper. You try to dissuade me from that choice by asking "what if everybody did that?" I can reply that none of us would be any worse off than we are now. A successful argument against my switching to paper requires a broader generalization principle. Any reasoning that makes it proper for me to switch to paper now will, after enough of us have made the switch, make it proper for others to switch to scissors, and the net result of all these changes will be harmful to everyone. This example should be of particular interest to utilitarians. If prisoner dilemma situations lead philosophers to abandon act utilitarianism in favor of utilitarian generalization, cyclic rock-paper-scissors situations should lead them to abandon common forms of utilitarian generalization.

3.5. Inappropriate Discrimination

The final class of examples is one that has been touted as a successful application of evolutionary game theory to ethics. I want to argue that it merits a closer look. Consider the Chicken game with the payoff matrix shown below

Chicken

	Hawk	Dove
Hawk	0,0	6,2
Dove	2,6	3,3

On the interpretation relevant for present purposes, Row and Column are equally strong players who would benefit equally by the possession of a single indivisible object. They can choose an aggressive strategy (hawk) or a submissive one (dove). When one chooses hawk while the other chooses dove, there is no fight and hawk immediately gets the object with probability one. If they both choose dove, there is no fight and each gets the object with probability one half (after a slight delay).[14] If they both choose hawk each has a one half chance of obtaining the object, but that chance for gain is outweighed by the fight required to obtain it.

It is not difficult to show[15] that the strategy that forms a nash equilibrium with itself in this game is the strategy of mixing hawk and dove in the ratio sixty-forty. Now consider the parallel evolutionary version of this game. To simplify discussion, assume that mixed strategies are not available and there is a continuum of players. (The n-player game can be understood as an approximation to this game.) Then the only profiles that could meet any of the solutions discussed are those in which 60% play hawk and 40% percent play dove. Under winner imitation, is highly unstable. If a single dove enters the population, the hawks will do slightly better than the doves, and so a fraction of the doves will become hawk. A single mutant dove will then convert a similar fraction of the hawks to doves and the population will cycle rapidly between mostly-hawk and mostly-dove.[16] Thus, there can be no universally stable populations in this game. The sixty-forty profile does, however, meets the MS condition. Under the replicator dynamic, therefore, the population will evolve to that state. Each player gets an average payoff of 2.2, which is even less than he would get if they all played dove.

Several authors take this game to illustrate Hobbes' dictum that a state of nature is a state of war. (Recall Hobbes' characterization that "..WARRE, consisteth not in Battell onely, or the act of fighting, but in a tract of time wherein the WILL to contend by Battell is sufficiently known;...''). For Hobbes, escape from this state of war is possible only with the emergence of a "sovereign" with power to enforce contracts. The more recent contention, which I want to reexamine, is that evolutionary game theory reveals a more plausible means of escape.

The clearest and most detailed expression of this idea is found in Sugden. Sugden argues that the chicken players may avoid their hawkish equilibrium if they become aware of an asymmetry that recurs in their meeting situations. Let us suppose, for example, that in some (non-zero) proportion of the meetings the desirable object is already in possession of one of the players. We call the resulting game *asymmetric chicken*. The one who has the object is *possessor* and the other player is *claimant*. Players can now consider strategies that are conditional on the role they find themselves playing. For example (h,d,h) is the strategy of playing hawk if possessor, dove if claimant, and hawk if neither. The information about a population needed to determine a player's average payoff can be given by a triple (p_1,p_2,p_3) where p_1, p_2 and p_3, are the proportions of the population playing hawk if possessor, claimant, and neither. Let us therefore extend previous usage and call such triples strategy profiles. Consider first the case of the *balanced* asymmetric game--each player has a fifty-fifty chance of playing the role of possessor in the asymmetric meetings. In that case the special symmetry assumption that characterizes our framework still holds. The payoffs to any player adopting the (i_1,i_2,i_3) conditional strategy against any player adopting the conditional (j_1,j_2,j_3) strategy are then the same, and so it makes sense to talk about the payoff to strategies rather than players. It is easy to show that (1,0,.6), (0,1,.6), and (.6,.6,.6) are the only profiles in the associated one-shot game that form nash equilibria with themselves, and that the first two of these satisfy condition MS. (Details are given in the appendix.) Call the strategies of the first set "retention" strategies–resources are used to try to retain what one has and those of the second set the "procurement" strategies–resources are used to try to procure what one lacks.

By noticing an asymmetry that is present in many of their meeting situations, the players are able to escape the hawkish equilibrium that appeared to have trapped them. It might seem a little embarrassing that there are now *two* equally stable equilibria, the one with retention and the one with procurement. There are a couple of considerations, however, that favor the former. First there is the psychological phenomenon that people place greater value on an object when it is in their possession than when it is not. This phenomenon has been confirmed experimentally on several occasions.(See [Knetsch and Sinden], Knetsch and [Samuelson and Zeckhauser].) Second there is the commonsense strategic observation that, in battle over a good, the odds of success and the cost of battle both favor the current possessor. Each of these factors would tend to change the payoffs of the chicken game in a way that increases retention's degree of stability and decreases procurement's. This may explain common conventions that allocate property to possessors. Possession, as Sugden reminds us, is nine points of the law.

All this reasoning was done under the assumption that each player has an equal chance of playing either role in his asymmetric meetings. A similar result can be obtained for the unbalanced game. Suppose that some players (called them "privileged") have a greater than 50% chance of being possessor in their asymmetrical meetings. Then there must be some others (call them "deprived") who have a less than 50% chance. Under these conditions our special symmetry assumption will be violated. A privileged player using retention will have a greater expected payoff against an opponent playing retention than a deprived player using retention would. A deprived player using procurement will have a greater expected payoff against an opponent playing procurement than a privileged player would. We argued above that when the special symmetry assumption is violated, there is little plausibility in employing the usual kinds of evolutionary dynamics and solution concepts, and we ought rather to consider dynamics in which each player compares his current payoff with those he received before. Let us examine how a population of uneven privilege fares under such a dynamic.

Suppose that the population starts from a position in which no asymmetry is recognized. As we have seen, they will evolve to the "state of war" within which hawks comprise 60% of the population. Now suppose a group of players begin notice the frequent asymmetry and begin to adopt, say, retention. The average payoffs to all the deprived players will begin to decrease since they are now facing more opponents (in the asymmetric situations) playing hawk. Any switch to hawk or procurement will leave these players even worse off, so they will tend to switch to dove or retention. The average payoffs to privileged players, on the other hand, will begin to increase since they are facing more doves than before. Any "accidental" switches to hawk or retention, however, will be rewarded and any accidental switches to dove or procurement will be punished. So the population will move towards one in which the privileged play hawk or retention and the deprived play dove or retention. If a privileged or deprived player is *maximally* privileged or deprived (so that he plays the same role in every asymmetric meeting), then there is no difference in the asymmetric meetings between playing hawk and retention or between playing dove and retention. Otherwise, any player under these conditions benefits by switching from his unconditional strategy to retention. As long as the dynamics permits the deprived to forget the higher average payoffs that they had gotten in the state of nature, we would expect retention to be stable. Marx's notion that members of the exploited class have "nothing to lose but their chains" turns out to be false in the short term. The deprived who decide not to adopt retention make things worse for themselves. The same argument, of course, shows that procurement is stable. As long as we don't adjust relative payoffs, the population will move towards whichever conditional strategy first appears within it. The appendix contains a more fastidious exposition characterizing all the nash equilibria and stable points of the unbalanced asymmetric chicken game.

The question to which I would like to draw attention is whether the most stable solutions to the unbalanced asymmetric property division games are the morally correct ones. One manifestation of this question concerns the appropriateness of the universal retention strategy when half are highly privileged and half are highly deprived. But the question arises even more forcefully when we consider that many other asymmetries are likely to be available to the players facing the good-acquisition problem. Suppose, for example, that all the players are either blue or green. Since a player can't choose his color, color is not "heritable" when the dynamics is understood as representing cultural evolution. Then half of all encounters will be between players of different colors. So the strategy "hawk if blue, dove if green" will be an evolutionarily stable strategy. This strategy clearly discriminates against greens. Any conflict between a blue and a green is settled in favor the blue. Yet, once the strategy becomes dominant, the greens must go along–to play hawk against a blue in this environment would be self-destructive.

The recognition of any asymmetry makes possible conventions that allow us to avoid Hobbes' "threat of battell," but some of these conventions seem to do so in an unfair way. What makes the convention that favors the possessor a "just" convention but the convention that favors the blue player unjust? We have seen that certain facts about psychology and strategy give the convention favoring possessors an especially high degree of stability. But this can't be what makes it right. For if the blues tended to be physically stronger than the greens (as they might if "blue" and "green" referred to males and females) then the policy that discriminates in favor of blues would get a similar boost in stability, with no concomitant boost in legitimacy. The problem of choosing which asymmetry to employ is, in game-theoretic terms, a coordination problem. It is frequently suggested that factors like salience and clarity determine solutions to such problems. It is difficult to contend, however, that the asymmetry of possession is more salient or clear-cut than the asymmetry of color. One might think that the asymmetry of possession is favored because it is more *pervasive.* We know, after all, that the color asymmetry is present in only 50% of the meetings in a hawk-dove game. Whether this observation is accurate depends on what we mean by a "possessor." If being in possession of something requires touching it, then it would seem that the possession asymmetry is present in far fewer than 50% of property-assignment games. No matter how pervasive the possession-asymmetry is, however, it can't be mere pervasiveness that makes possessor-discrimination right and color-discrimination wrong. A convention based on color-asymmetry would be wrong even if it applied only to cases in which there was no possession-asymmetry. And a convention based on, say, foot length asymmetry, would be wrong even if such asymmetry were present in 100% of meetings.

The existence of equilibria in hawk-dove games based on inappropriate discrimination is another example that deserves attention from those who wish to apply ethics to game theory. That is not to say there is anything fallacious about the game theoretic reasoning itself. Stable patterns involving similar discriminations are quite familiar, whether it be by sex among Taliban in Afghanistan or by antler size among moose in Maine. The question is why, as long as initial distributions are sufficiently equal, discrimination based on possession is regarded as morally proper, while discrimination based on color is not.

One possible answer to this question appeals to the same ideas as the earlier discussion of suboptimal equilibria. Morality is an institution whereby psychological rewards and punishments can change the payoffs in a game. This change can create an equilibrium where none existed before (as in the prisoner's dilemma example) or it can buttress (i.e., make more stable) an already existing equilibrium. What makes morality possible is our peculiar ability to make each other feel guilt and pride at little or no cost to ourselves. It is not true, of course, that one person can cause any other to feel guilt or pride whenever he wishes. It is rather that, by an appropriate moral education, we can collectively program ourselves to feel guilt or pride whenever we become aware that certain general conditions obtain.[17] Furthermore, we are aware of this ability we have and of its usefulness. We all campaign for various curricula that might guide our moral education. Just as in electoral campaigns, our support depends both on how we and those we care about would fare were the curriculum adopted and on the prospects that others will support it.

Consider now how players in the hawkish equilibrium in which no asymmetry is recognized might evaluate curricula that inculcate conventions based on color and on possession. More specifically, let us suppose that goods are distributed so that half of all meetings involve a possession asymmetry and each player is possessor in half of his asymmetric meetings. Then a possession convention nets each player 2.2 units in symmetric meetings and 4 units in asymmetric meetings, for an overall expectation of 3.1 units. This represents a considerable improvement over the 2.2 units expected in the state of nature equilibrium. Under a color-based convention, the favored players will expect 2.2 units in symmetric meetings and six units in asymmetric meetings, for an overall expectation of 4.1 units. The unfavored players, however will receive only 2 units in asymmetric meetings and their overall expectation will therefore be only 2.1 units, which is less than they get in the state of nature. Thus every player might support, and expect all others to support, a curriculum to inculcate a convention based on possession. Nobody could expect similar support for a curriculum to inculcate a convention based on color.

The results of this line of thinking are similar in some ways to the results of John Rawls' choice behind a veil of ignorance. For Rawls inequalities are permissible if and only if they result in everybody's being better off than they are in a state of equality. In our hawk-dove game the state of nature is a state in which we are all equally miserable. Inequalities are permissible only if nobody is worse off than in the state of nature. For Rawls, the device that ensures the elimination of the bad inequalities is the veil of ignorance. Since I might be disadvantaged myself, I cannot support an institution that makes the disadvantaged worse off. Here the players have full awareness of their circumstances and inequalities are limited by the need for consensus. It is because I know that *others* would be disadvantaged that I know a proposed curriculum will not get the necessary support.

The campaign and election metaphors suggest a retrenchment from the evolutionary framework within which this discussion has thus far been conducted. It may be possible to resist this suggestion. The selection of a component of moral education could itself be viewed as an evolutionary game that accompanies and modifies the game modeling the behavior it aims to regulate. The result would be a compound game played as follows. A strategy for each player is a move (possibly mixed) in both the behavioral game and the regulatory game. A move in the regulatory game consists of a decision to advocate the color asymmetry, the possession asymmetry or neither. In each round every pair of players plays the behavioral game and each player also makes a move in the regulatory game. Strategies are updated by the appropriate dynamic, with payoffs computed from the behavioral game in the usual way. If sufficiently many players make the same move in the regulatory game, payoffs are updated appropriately before the next round.

It is not obvious that an evolutionary view of regulation can account for the superiority of retention to color discrimination as successfully as the picture of norms chosen at one shot by agents who understand the effects on themselves and others of all candidates. If we rule out *hypocrisy,* so that advocacy of a convention requires adherence to it, then advocating an equilibrium different than one currently practiced will be costly. Potential reformers will be driven extinct. Even if we don't rule out hypocrisy, it is difficult to see how the reformer benefits by his advocacy. There are, however, several considerations that may provide hope for an evolutionary account. Advocacy is typically a *public* action, so a player's moves in the behavioral game might be conditioned on her opponent's moves in the advocacy game.

Furthermore, advocacy often causes others to e*xpect* that the advocate will adhere to the convention he advocates. As Sugden has observed, it is part of our psychological constitution to suffer when others are hurt by false expectations about our behavior. So it may turn out that success comes from non-hypocritical advocacy and behavior that assumes non-hypocritical advocacy on the part of one's opponent. Furthermore, once a position is widely advocated, the advocacy itself is probably at least weakly rewarded. As long as there are no countervailing factors, this may be enough to push a common move in the regulatory game across the threshold of efficacy. I hope to explore such possibilities in future work.

Appendix: Proofs of Claims.

1.For G an evolutionary game with a continuum of players and a strategy profile of G, let G_n be the n-player version of G and $_n$ be a closest approximation to in G_n. There are games G with an equilibrium such that: a) the replicator dynamics carries $_n$ to and is stable, b)replicator carries $_n$ to and is not stable, and c) replicator does not carry $_n$ to .

Proof. Consider the following two-player game matrices.

	A	B
A	0,0	1,1
B	1,1	0,0

	A	B	C
A	0,0	1,1	0,1
B	1,1	0,0	0,1
C	1,0	1,0	2,2

	A	B
A	1,1	0,0
B	0,0	1,1

Let G^1, G^2, G^3 be parallel evolutionary games with a continuum of players under the replicator dynamics based on these matrices and let G^1_{101}, G^2_{101},G^3_{101} be versions of these games with 101 players. Then σ=(.5,.5) is a stable equilibrium of G^1 and an unstable equilibrium of G^3. σ_{101}=(50/101,51/101) is a closest approximation to σ. In G^1 replicator carries σ_{101} to σ. In G^3 replicator carries σ_{101} away from σ to the more distant equilibrium (0,1). In G^2, replicator carries (50/101,51/101,0) to the unstable equilibrium (.5,.5,0).

2.There are parallel evolutionary games with strategies meeting condition A but: neither MS nor BL, MS but not BL, and BL but not MS.

Proof. Consider the following two-player game matrices:

	A	B
A	0,0	0,0
B	0,0	1,1

	A	B	C
A	1,1	1,1	1,1
B	1,1	0,0	2,2
C	1,1	2,2	0,0

	A	B
A	1,1	0,1
B	1,0	2,2

Strategy A is a weak nash equilibrium in all three games. In the first one it satisfies neither MS nor BL, in the second it satisfies MS but not BL, and in the third it satisfies BL but not MS.

1. Let G be a two person game with payoffs V_G(i,j) for moves i,j {1,…,k}. (,) is a nash equilibrium of G if and only if V(,) V(j,) for all pure strategies j {1,…,k}.

Proof. Suppose that (,) is a nash equilibrium. Then V(,) V(,) for all strategies . In particular this holds whenever one of the weights in is one and the others are all zero, i.e., V(,) V(j,) for all j {1,…,k} as was to be shown. Conversely, suppose V(,) V(j,) for all j {1,…,k}. Since, for any , V(,) is a mix of these V(j,), V(,) V(,) for any strategy , as was to be shown.

2.Let G be as above and let R be a “round-robin” version of this game with a continuum of players. Each player chooses a pure strategy i and plays G with each of the others one time and gets a payoff V_R (i,) equal to the average of the payoffs from each encounter. Note that =(S_1,…,S_k) can be regarded as either a mixed strategy in G or a profile in R. Since there are a continuum of players, a single encounter has no effect on the average, and so V_R (i,)= V_G (i,).

a. (,) a nash equilibrium of G implies a nash equilibrium of R.

Proof. Suppose =(S_1,…,S_k) and (,) a nash equilibrium of G. Let J be the set of all i {1,…,k} such that S_i 0. Then, for all j, in J, V_G(j,)=V_G(,), for otherwise we could adjust the weights S_1,…,S_k in to obtain a for which V_G(,)>V_G(,). Suppose is not nash in R. Then there is some player playing strategy j who could benefit by unilaterally switching to j , i.e., V_R(j ,)>V_R(j,). Since j J and V_R=V_G, this implies V(j ,)>V(,), which, by 1 is impossible.

b. Then a nash equilibrium of R implies (,) a nash equilibrium of G.

Proof. Suppose =(S_1,…,S_k) is a nash equilibrium of R. Let J be the set of all i {1,…,k} such that si 0. Then, for all j,j in J, V(j,)=V(j ,), for otherwise the strategy that did worse would benefit from switching. Since is composed of members of J, it follows that V(j,)=V(,) for j J. For j J and j’ J, V(j,) V(j ,), for otherwise j would benefit from switching to j’. Hence V(,) V(j,) for all pure strategies j, j {1,…,k}, and so by part 1 above, (,) is a nash equilibrium of G

3)In the balanced asymmetric chicken game each of the strategies (1,0,.6), (0,1,.6), and (.6,.6,.6) forms a nash equilibrium when paired with itself.

Proof. Consider an arbitrary mixed strategy p=(p_1,p_2,p_3), of playing hawk with probability p_1 as possessor, p_2, as claimant and p_3 as neither. V(p, (1,0,.6)) = $.5\bar{r}(6p_1+3\bar{p}_1)+.5\bar{r}(2\bar{p}_2)+r(6p_3.4+2\bar{p}_3.6+3\bar{p}_3.4)$ where r is the probability of an asymmetric meeting.

By arithmetic, $V(p, (1,0,.6)=\bar{r}(1.5p_1+\bar{p}_2+1.5)+r(2.4)$. This expression reaches its maximum value whenever $p_1=1$ and $p_2=0$, and so (1,0,.6) is a best reply to itself. A similar argument establishes that ((0,1,.6), (0,1,.6)) is a nash equilibrium. Now consider (.6,.6,.6). Since (.6,.6) is a mixed equilibrium in the symmetric chicken game, V(s,.6) has the same value v for every strategy s in that game. Since V(p,(.6,.6.,.6)) is a mix of $V(p_1,.6)$, $V(p_2,.6)$ and $V(p_3.6)$, it follows that V(p,(.6,.6,.6)) =v for every conditional strategy p. Thus (.6,.6,.6) is a best reply to itself.

4) In the balanced asymmetric chicken game no strategies other than those mentioned above form a nash equilibrium when paired with themselves.

Proof. Suppose that $p=(p_1,p_2,p_3)$, were another. If $p_3<.6$ then $(p_1,p_2,1)$ would be a better reply to p than p itself. Similarly if $p_3>.6$, then $(p_1,p_2,0)$ would be a better reply to p. Hence $p_3=.6$. Now we examine p_1 and p_2. Because p is not among the strategies mentioned in 1, at least one of these strategies must be different than .6. Without loss of generality we may suppose it is p_1. If $p_1<.6$, then any best reply to p has 1 for its second component, so $p=(p_1,1,.6)$. Any best reply to this has 0 as its first component, so p=(0,1,.6) which was listed in 1 after all. By similar reasoning, if $p_1>.6$ then p_2 must be 0 and p_1 must be 1, so p=(1,0,.6) which, again, was listed in 1. So the supposition that p was not listed is false and there are no more strategies that form nash equilibria with themselves

5)In the balanced asymmetric chicken game the profiles represented by (1,0,.6) and (0,1,.6) satisfy condition MS, but the profile represented by (.6,.6,.6) does not.

Proof. Let   be the (unique) profile represented by (1,0,.6). Take any (pure conditional) strategy $j=(j_1,j_2,j_3)$. If j_1 h or j_2 d then V(,)>V(j,). If j_1=h and j_2=d, then V(,)=V(j,), but (since V(.6,h)>V(h,h) and V(.6,d)>V(d,d) in the original hawk-dove game) V(,j)>V(j,j). The proof for (0, 1,.6) is similar.   To see that (.6,.6,.6) fails to satisfy MC, note that, because every strategy scores 2.4 against .6 in the chicken game, V((.6,.6,.6),(.6,.6,.6))=V((.6,.6,.6), (1,0,.6)). V((1,0, .6),(1,0,.6)) is a mix of 2.4 with a fifty-fifty mix of six and 2, and so it is greater than

V((.6,.6,.6),(.6,.6,.6)).

6)In the unbalanced asymmetric chicken game the profiles (1,0,.6), (0,1,.6) and (.6,.6,.6) are generalized nash equilibria.

Proof. In the unbalanced game we can no longer identify a player j with her strategy triple (j_1,j_2,j_3) because her expected payoff now depends on whether she is privileged or deprived as well as the strategy she employs. So we adopt the notation $j=(I,j_1,j_2,j_3)$ where I=P (for "privileged" or D for "deprived") to denote a player j with status I and strategy triple (j_1,j_2,j_3). Let $j=(I,j_1,j_2,j_3)$ be any member of the population with profile   =(1,0,.6). We write V(j,) for the payoff to j in the population with profile . (Note that the status of j and the strategies of j and are sufficient to determine this payoff.) It is required to show that, for every strategy j , V(j,) V(j ,). V(j,)=r(p6+$\bar{p}$2)+$\bar{r}$.6 where p is the probability that j is possessor in the asymmetric meetings and r is the probability of an asymmetric meeting. Note that the value of p (and hence that of V(j,)) depends on whether j is privileged or deprived. Now take any j j such that j has the same status as j. If j differs from j only in the fourth coordinate then V(j ,)=V(j,). Otherwise, either j 1<1 and j $_2$ 0 or j 1 1 and j $_2$>0. Hence V(j ,)=r(ph+$\bar{p}$k)+$\bar{r}$.6, where either h<6 and k 2 or h 6 and k<2. Either way, V(j ,) < V(j,), as was to be shown. The proof for (0,1,.6) is similar. The claim that (.6,.6,.6) is nash follows immediately from the observation that, for all strategies j and j , V(j,(.6,.6,.6))=V(j ,(.6,.6,.6).

7)There are no nash equilibria in the unbalanced asymmetric chicken game other than those mentioned above.

Proof. Suppose that there were another represented by (p_1,p_2,p_3). If p_3<.6, then any member $j=(I,j_1,j_2,j_3)$ of the population with j_3=d could improve his payoff by switching to (I, j_1,j_2,h), and if p_3>.6, then any member $j=(I,j_1,j_2,j_3)$ with j_3=h could improve her payoff by switching to (I,j_1,j_2,d). Hence p_3=.6. Now suppose p_1=0, i.e., no players in the population play hawk when possessor. Then any players with a strategy of the form (I,j_1,d,j_3) would benefit by switching to

(I,j_1,h,j_3), so there can be no such players and $p_2=1$. But (0,1,.6) was already listed so p_1 0. Similar arguments establish that p_1 1, p_2 1 and p_2 0. Because (p_1,p_2,p_3) is not among the profiles mentioned above, either p_1 or p_2 must be different than .6. Without loss of generality we may suppose it is p_1. If $p_1<.6$, then, any player with a strategy of the form (I,j_1,d,j_3) would benefit by switching to (I,j_1,h,j_3). There must be such a player because $p_2<1$. If $p_1>.6$ then any player with a strategy of the form (I,j_1,h,j_3) would benefit by switching to (I,j_1,d,j_3). There must be such a player because $p_2>0$. We have now shown that p_1 is neither equal to .6 greater than .6 or less than .6. Hence the supposition is false and there are no other equilibria.

Notes

[1] Here I have in mind Braithwaite and, more recently, John Rawls, David Gauthier, Brian Skyrms, Peter Danielson, and Peter Vanderschraaf among others.

[2] Here I have in mind John Harsanyi, Robert Sugden, Kenneth Binmore, Jonathan Bendor and Piotr Swistak among others.

[3] It makes sense, that is, as long as attention is restricted to *asexual* reproduction. Even for biological applications, the replicator dynamics has important limitations.

[4] See Börgers and Sarin. The learning model investigated there originates in Bush and Mosteller and Cross. In Börgers and Sarin the payoffs are assumed to lie strictly between 0 and 1 and *absolute* payoffs are used to compute the new mix. It is shown that under these conditions that the expected weights in each player's mix evolve according to replicator dynamics. Fascinating as this observation is, it provides no justification for assuming that the replicator dynamics describes the evolution of the proportions of the population playing particular strategies in games modeling cultural evolution. It is not clear how to characterize the evolution of a population of mixers, each of which itself evolves by replicator dynamics.

[5] For example, that if the population has n members for odd n and replicator calls for a profile ((n-1)/2n, (n+1)/2n) to be replaced by the unattainable (.5,.5) then the population remains as it was, and that if a dynamics causes a population to cycle between "adjoining" profiles ((n-1)/2n, (n+1)/2n), and ((n+1)/2n, (n-1)/2n), then these profiles be considered equilibria.

[6] The observation that one can show that these dynamics arise under various stories about the procedures by which players actually change strategies does render them independent of interpersonal comparisons. In the Bendor and Swistak characterization of the replicator

dynamics, for example, each player must know his own *relative* payoff V_i/V, which implies that his payoff is comparable to those of the others. This is to be expected, since the replicator dynamics itself requires such comparisons. A strategy whose frequency is decreasing under this dynamic might well begin to increase if the payoffs to a player employing it were to be doubled relative to the others.

[7] For example, Sugden (pp14-19) writes "Utility indices, as I shall interpret them, are intended only to represent each individual's preferences over sequences of outcomes; they are not to be understood as conveying any judgements about the relative intensity of different individuals' wants." It is not clear what dynamics Sugden has mind, but it does seem plausible that he is construing his solution conditions (A and MS) counterfactually.

[8] In particular, a player who is maximizing expected payoff will hunt hare only if 3p>2, where p is the probability that she assigns to the other hunting hare, and so the preferred equilibrium will be attained only if each assigns a probability of at least 2/3 to the other hunting hare.

[9] The calculation is similar to the previous one. The payoff to a native hare hunter in a population with proportion p of stag hunters exceeds the payoff to the invaders as long as 2>3p, i.e., as long as p<2/3.

[10] They are those that are widely followed in a group and that have the property that a follower always benefits from others' following.

[11] The distinction between appraiser and agent relativism is from Lyons

[12] The condition is $(T_1-R_1)(T_2-R_2) < (R_1-S_1)(R_2-S_2)$, where Ti, Ri, and Si are the reward, sucker and punishment payoffs for player i. The condition is discussed in [Kuhn and Moresi] and Kuhn.

[13] A referee pointed out that such appeals to randomizing devices do seem to be found in the scriptures. There are several biblical references, for example, to "the Urim and the Thummim" which appear to be two small, similarly shaped stones that could be placed in the breastpiece of a priest, and which could be "consulted," perhaps by drawing one blindly from the pair. Such consultation was used to determine the guilt or innocence of an individual and to make some other binary moral decisions. Similarly, there are many biblical references to making important moral decisions by casting of lots. In both cases, however, the decision making apparatus seems to be regarded, not as a device for randomizing, but as a method of learning God's will. For example, in Pr 16:33 we have "The lot is cast into the lap but the decision is wholly from the Lord." The choice is determined even though it may appear random to us. A better counterexample from more recent literature can be found Taurek, where the suggestion is made that if I must choose between saving the inhabitants of island A or the inhabitants of island B (all of whom are strangers to me), I ought to base my decision on the flip of a coin (regardless of the populations of A and B), thus affording every person an equal chance of survival. I suspect that utter implausibility of this conclusion (and the skill with which Taurek defended it) was part of

the reason that the paper has received so much attention.

[14] If the value of the object is four and the value of avoiding the fight is two, the value of avoiding the delay is one. Nothing important in what follows would change if we drop the assumption that there is a delay in dove-dove interactions over dove-hawk interactions, but the equilibrium could be less "hawkish" and the state of nature equilibrium could produce payoffs greater than the dove-dove payoffs. Thus the delay makes the situation more dramatic. Furthermore, it is not unrealistic. The mechanism by which the good is "distributed" is much more obvious in the case of dove-hawk interactions than in the case of "dove-dove" interactions.

[15] Since the best reply to hawk is dove and the best reply to dove is hawk no pure strategies can meet this condition. At the mixed equilibrium where hawk is played with probability p, hawk and dove must get equal payoffs. So $0p+6\bar{p}=2p+3\bar{p}$, and so p=.6, as was to be shown.

[16] A similar argument applies to a population   of players who mix at 60-40. Suppose a single mutant d playing dove invades. Since   is nash V(d,)=V(,), and the mutant survives under winner imitation. Now suppose a single mutant hawk h invades. Since V(h,d)>V(,), the invader will do slightly better than the natives and many doves will switch to hawk. Now a single mutant dove will convert many hawks to doves and the population will cycle as above. All this suggests that winner imitation leads to instability when there are mixed strategy nash equilibria in the underlying game.

[17] The locutions "feeling guilt or pride" suggests that moral education causes people to experience certain feelings *in addition to* what they would have experienced without the education. Sometimes it seems instead to change the nature of our experience altogether. Children may come to feel sharing pleasurable rather than irksome and bullying tiresome rather than exciting. For the purposes of this discussion all that matters is that we have the ability to collectively "program" ourselves so as to change the payoffs in the game situations we will face in the future.

Acknowledgments

This paper has benefitted from my discussions with Brian Skyrms, Piotr Swistak, and the members of the evolutionary game theory seminar at UC Irvine and from the sabbatical support of Georgetown University.

References

Bendor, J. and P. Swistak:1997, 'The Evolutionary Stability of Cooperation', *American Political Science Review* **91: 2**, 290-307.

Bendor, J. and P. Swistak:1998, 'Evolutionary Equilibria: Characterization Theorems and Their Implications', *Theory and Decision*, **45:2**, 99-159.

Binmore, K.G.: 1998 *Game Theory and the Social Contract. Vol 1*(1994) *Playing Fair*, Vol 2 (1998) *Just Playing.* Cambridge, MA: MIT Press

Boyd, R. and J. Lorberbaum: 1987, 'No Pure Strategy is Evolutionarily Stable in the Repeated Prisoner's Dilemma Game', *Nature*, **327** (7 May), 58-59.

Börgers, T. and R. Sarin: 1997, 'Learning Through Reinforcement and Replicator Dynamics', *Journal of Economic Theory* **77**, 1-14.

Bush, R. and F. Mosteller, 1951, 'Stochastic Models For Learning', *Psychological Review* **58,** 313-323

Cross, J.: 1973, 'A Stochastic Learning Model of Economic Behavior', *Quarterly Journal of Economics* **87**, 239-266.

Gauthier, D.:1967, 'Morality and Advantage', *Philosophical Review* **76:4**, 460-475.

Gintis, H.: 2000, *Game Theory Evolving*, Princeton University Press.

Knetsch, J.L.:1989, 'The Endowment Effect and Evidence of Non-Reversible Indifference Curves', *American Economic Review* **79** , 1277-1284.

Kuhn, Steven and Serge Moresi, "Pure and Utilitarian Prisoner's Dilemmas," *Economics and Philosophy* **11** (1995) pp 123-133

Kuhn, Steven "Agreement Keeping and Indirect Moral Theory," *Journal of Philosophy*, **93** (1996) pp105-128.

Knetsch, J.L. and J.A. Sinden: 1984, 'Willingness to Pay and Compensation Demanded:

Experimental Evidence of an Unexpected Disparity in Measures of Value," *Quarterly Journal of Economics* **99**, 131-139.

Lyons, J.:1976, 'Ethical Relativism and the Problem of Incoherence', *Ethics* **86,** 107-121.

Nowak, M. and R. May,:1992, 'Evolutionary Games and Spatial Chaos', *Nature*, **359**, 826-829.

Nowak M A, S. Bonhoeffer and R. May,:1994, 'Spatial Games and the Maintenance of Cooperation', *Proceedings of the National Academy of Sciences* **91**, 4877-4881.

Samuelson and Zeckhauser, 1988, 'Status Quo Bias in Decision Making', *Journal of Risk and Uncertainty* **1,** 7-59.

Selten, R.:1983, 'Evolutionary Stability in Extensive Two-person Games', *Mathematical Social Sciences*, **5**, 269-363.

Sen, A.: 1967, 'Isolation, Assurance and the Social Rate of Discount', *Quarterly Journal of Economics* **81:1**, 112-124.

Skyrms, B.: 2001, 'The Stag Hunt', *Proceedings and Addresses of the American Philosophical Association* **55:2**, 31-41.

Skyrms, B.: 1999, 'Reply to Critics', *Philosophy and Phenomenological Research* **59:1,**

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

Skyrms, B.; 1994, 'Darwin Meets 'The Logic of Decision', *Philosophy of Science* **61**, 503-528.

Skyrms, B. and J. Alexander: 1999, 'Bargaining with Neighbors: Is Justice Contagious?' *Journal of Philosophy*, 588-598.

Skyrms, B. and R. Pemantle: 2000, 'A Dynamic Model of Social Network Formation', *Proceedings of the National Academy of Sciences* **97:16** , 9340-9346.

Sugden, R.: 1986, *The Economics of Rights Cooperation and Welfare*, Oxford: Basil Blackwell.

Taurek, J.: 19 , *Philosophy and Public Affairs* **2:1**, 293-316.