
THE JOURNAL OF PHILOSOPHY

VOLUME CII, NO. 3, MARCH 2005

A PLEA FOR ASYMMETRIC GAMES*

Game theory has been an important tool for addressing difficult problems regarding the origins of conventions, fairness, and prosocial behaviors in general. Examples of the successful use of game theory include David Lewis's seminal work on conventions,¹ John Nash's theory of equilibrium behavior in noncooperative games,² and Robert Axelrod's use of the iterated Prisoner's Dilemma to model the evolution of cooperation.³

One feature of game-theoretic models that makes them so useful is their simplicity. However, simple formal models of complex phenomena are a double-edged sword. On the one hand, when an explanation is convincing, and that explanation uses only a simple model of the explanandum, it may possess several important explanatory virtues. For instance, its simplicity may indicate that a variety of features of the phenomenon may be irrelevant or at least unimportant. If so, then we may begin justifiably to suspect that the explanation may apply to other domains. As an example, the same formal tools have been applied successfully to both social learning processes and biological evolutionary processes.

But on the other hand, simple formal models omit details that may turn out to be important. If the devil is in the details, then simple formal models will not provide the robust explanations that we seek. Accordingly, when a simple model has been successfully applied to

* Portions of this paper were presented at Texas Tech University, Florida State University, and the California Institute of Technology. I would like to thank the faculty of those departments for valuable discussion. Thanks are also due to Jim Joyce, Sara Rachel Chant, and Michael Ruse.

¹ *Convention: A Philosophical Study* (Cambridge: Harvard, 1969).

² "The Bargaining Problem," *Econometrica*, XVIII (1950): 155–62.

³ *The Evolution of Cooperation* (New York: Basic Books, 1984).

one domain, there is the temptation to apply that technique to another domain for which it may not be appropriate.

Noncooperative game theory is such a simple and powerful tool that this temptation is particularly strong. The basic tenets of game theory can be stated using nothing more than high-school algebra, and even evolutionary game theory can be explained using mathematics no more complex than that found in a first-year calculus course. In spite of this simplicity, it has been deployed to analyze evolutionarily evolved behaviors, social learning processes, the formation of conventions, and the behavior of economic markets, to name just a few of its applications.

It is just a short step from these applications to the task of explaining the origins of our moral practices. After all, our moral practices resemble conventions, and the social contract may be thought of as a set of self-enforcing equilibrium behaviors. Thus, when we observe a practice of following principles such as “share and share alike,” a tendency to cooperate in Prisoner’s Dilemmas, or to follow broadly utilitarian principles, we may ask whether noncooperative game theory might explain these tendencies.

In this paper, I explore the use of game theory in explaining the origins of our moral practices. Negatively, I will argue that the great generality of game-theoretic explanations has encouraged philosophers to misconstrue the project in such a way as to block progress and invite a raft of criticisms. But positively, I will argue that such excesses are not a necessary component of the project. Specifically, I will argue that the generality of game-theoretic explanations may be abandoned in favor of a legitimate pluralism. That is, game theory may be used to construct a variety of specific explanations of particular moral practices, rather than a single explanation of diverse moral practices.

I. GAMES AND BLACK BOXES

Individuals often cooperate in one-shot Prisoner’s Dilemmas, share windfalls even when they are not forced to do so, form coalitions, and follow conventions even when there is no penalty for deviation. Such behaviors require explanation, because the standard *Homo economicus* conception of human beings as perfectly rational, self-interested welfare maximizers predicts that such prosocial behaviors would not be observed nearly as often as they actually are.

There is a simple procedure for using game theory as the primary explanatory tool for explaining the origins of prosocial behaviors. First, we strip away the irrelevant details of the situation, and focus only on the possible actions of the agents, as well as the payoffs the agents will obtain, contingent upon their actions. For instance,

experimental economists have performed experiments in which individuals must determine how to split a \$10 windfall between them. Not surprisingly, they overwhelmingly favor a 50/50 split, even when the participants cannot communicate with each other. The interesting question here is not “What do the agents do?” for it is obvious that they will split it evenly. Rather, the interesting question is “Why is it so obvious that they will favor the 50/50 split?” In other words, we want to know if there is anything about a 50/50 split that accounts for why it is such an intuitively compelling and expected outcome.

A game theorist such as Brian Skyrms (who has examined this game in detail⁴) will begin by simplifying the situation. From a game-theoretic perspective, it is not important that the subjects are participating in an experiment, that they are splitting money rather than some other divisible good, what their environment is like when they are negotiating, and so on. The only important questions that we consider are “What are the possible actions available to each participant?” and “What are the possible payoffs that accompany those actions?”

Those two questions lead directly to the game-theoretic model. We may suppose (following Skyrms) that each participant may demand either one-third, one-half, or two-thirds of the money. We may settle the second question by supposing that if their demands add up to 100% or less of the money, then each participant gets what she demanded, but if their demands are greater than 100%, then each gets nothing.

Not surprisingly, there are a variety of considerations suggesting that those two facts by themselves are sufficient to favor the 50/50 split. Although those arguments have been questioned in other contexts,⁵ I will delay considering those complications until later. What is important at this point is that this game-theoretic explanation omits the mechanisms that are supposed to give rise to the “fair” behavior. For instance, it is possible that the participants’ behaviors are the result of social learning processes—perhaps each participant has had the opportunity to observe others in similar situations, and is following the behavior that has led to the best outcomes in those cases. Or perhaps each is hypothetically considering the various alternatives

⁴ “Sex and Justice,” this JOURNAL, XCI, 6 (June 1994): 305–20; *Evolution of the Social Contract* (New York: Cambridge, 1996).

⁵ Justin D’Arms, “Sex, Fairness, and the Theory of Games,” this JOURNAL, XCVI, 12 (December 1996): 615–27; D’Arms, Robert Batterman, and Krzysztof Gorny, “Game Theoretic Explanations and the Evolution of Justice,” *Philosophy of Science*, LXV (1998): 76–102; Martin Barrett, Ellery Eells, Branden Fitelson, and Elliott Sober, “Models and Reality: A Review of Brian Skyrms’s *Evolution of the Social Contract*,” *Philosophy and Phenomenological Research*, LIX (1999): 237–42.

and has hit upon a convincing argument in favor of the 50/50 split.⁶ On the other end of the spectrum, perhaps the participants' behavior is the result of a hard-wired behavioral tendency that is the result of evolution and natural selection.

As it stands, the game-theoretic explanation treats all of these mechanisms as black boxes. Instead of focusing on the particular mechanisms that may be the proximate causes of the observed behavior, the game theorist assumes as a working hypothesis that there is something about the available strategies and payoffs themselves that is sufficient to explain and predict the observed behavior. If so, then any causal process that operates in such a situation will favor the 50/50 split, even if we do not specify whether the mechanism is social or biological. Accordingly, we have Skyrms's representative comment that social and biological processes may be treated analogously:

These biological concepts also have qualitative analogues in the realm of cultural evolution. Mutation corresponds to spontaneous trial of new behaviors. Recombination of complex thoughts and strategies is a source of novelty in culture. Using these tools of evolutionary dynamics, we can now study aspects of the social contract from a new perspective.⁷

If Skyrms is correct, then game-theoretic explanations of prosocial behaviors may be extremely powerful. Although it tempting to charge that Skyrms has simply disregarded the relevant differences between biological and social processes, this objection oversimplifies his position. According to game theorists, the real explanatory burden is to be placed upon the payoffs and stability properties of the interaction itself, and not on the proximate causes of behavior. This is not an unmotivated position, for the ability to subsume diverse phenomena under a single explanatory argument pattern is often taken to be the hallmark of explanation.⁸

Furthermore, there is another important reason why game theorists disregard the specific mechanisms that may play a role in the evolution of prosocial behaviors. Specifically, the mechanisms that typically accompany game-theoretic explanation tend to converge on the same predictions for a variety of games. And as a general point, if a variety of proximate mechanisms can all be shown to lead to the same observed phenomena, then one can argue that those mechanisms may be omit-

⁶ As in Nash's original conception of the bargaining process.

⁷ Skyrms, *Evolution of the Social Contract*, p. x.

⁸ See Philip Kitcher, "Explanatory Unification," in Joseph Pitt, ed., *Theories of Explanation* (New York: Oxford 1988), pp. 167–87.

	Stag	Hare
Stag	7,7	0,3
Hare	3,0	2,2

Figure 1: The Stag Hunt. Payoffs to the row player are first; payoffs to the column player are second.

ted from explanations.⁹ Indeed, this seems to be the case regarding the three types of proximate mechanism that are usually invoked in evolutionary explanations of prosocial behavior: equilibrium, assortment, and efficiency. We may illustrate each in turn.

1.1. Equilibrium. The type of explanation that was pioneered by Nash (*op. cit.*) in the 1950s does not advert to the causal history of a population of individuals, and it deliberately sets aside the issue of how the individuals in a population are supposed to converge on a prosocial behavior. Rather, equilibrium explanations make use of the fact that some collective behaviors are more resistant to change than others.

For example, consider the Stag Hunt game. This game is used to model a situation in which two players can either cooperate to achieve a high payoff, or go on their own to achieve a low, but guaranteed payoff. The game is represented in Figure 1, and is traditionally accompanied by the following story. There are two hunters, who each face a choice between hunting a stag and hunting a hare. It takes two hunters working together to hunt successfully a stag, so they will only achieve the higher payoff if both cooperate by hunting stag. However, each may strike out on her own and hunt a hare, for a low, but guaranteed, payoff.

The Stag Hunt provides a simple illustration of equilibrium explanations in game theory. Consider a population in which half the individuals play Stag, and the other half play Hare. Furthermore, suppose that individuals pair up randomly, play the game with each other once, and then disperse back into the population. The Stag players will receive a payoff of 7 half the time, and a payoff of 0 the other half, for an expected payoff of $3\frac{1}{2}$. Hare players will receive a payoff of 3 half of the time and a payoff of 2 the other half of the time. Thus, Hare players will receive an expected payoff of $3/2$, which is less than the Stag payoff.

⁹ Compare with Sober's notion of an "equilibrium explanation" in "Equilibrium Explanation," *Philosophical Studies*, XLIII (1983): 201–10.

Thus, under reasonable assumptions, we can expect this state of affairs to be unstable. If we are working within a social learning context and the players seek to maximize their own payoffs, then those who play the Hare strategy have an incentive to switch strategies. If we are thinking in terms of biological evolution and natural selection, then Stag players will have a higher average fitness than Hare players. And under reasonable assumptions, we would expect the population to approach a state in which everyone plays Stag.

On the other hand, a state in which everyone plays Stag is a stable equilibrium. In such a population, each player gets a payoff of 7 in every interaction. Thus, if someone were to switch to a strategy of playing Hare, her payoff would go down from 7 to 3. Because no individual has an incentive to switch strategies in a population where everyone plays Stag, we say that the Stag strategy is a Nash equilibrium.

Similarly, the state in which everyone plays Hare is also a Nash equilibrium. For in such a population, anyone who switches from Hare to Stag will lower her payoff from 2 to 0. However, here the notion of a stochastically stable equilibrium is useful.¹⁰ If the population is subject to random shocks, it is easier for the population to be moved from the Hare equilibrium to Stag equilibrium than vice-versa. Thus, we say that the Stag equilibrium is the unique stochastically stable Nash equilibrium, and is therefore the one we should expect the population to spend most of the time in over the long run.

I.2. Assortation. Along with stability properties, mechanisms of assortment have justifiably taken center stage in game-theoretic explanations of prosocial behavior. An assortment mechanism is any mechanism that raises the probability that similarly behaving individuals will interact with each other. These mechanisms may take the form of a tendency to interact with genetic relatives with similar genetically programmed behavioral tendencies, or a tendency deliberately and rationally to seek out like-minded individuals in a social context.

If there is such an assortative mechanism, then the expected payoff of the Stag strategy will be significantly higher than the expected payoff of the Hare strategy. For when two Stag players interact, each receives a payoff of 7, whereas when two Hare players interact, each receives a payoff of only 2. So if the assortment mechanism operates to a sufficiently high degree, the Stag strategy is the one we would expect to observe over the long run.

¹⁰ For a clear exposition of the concept of a stochastic stability, see H. Peyton Young, *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions* (Princeton: University Press, 1998). See also Larry Samuelson, *Evolutionary Games and Equilibrium Selection* (Cambridge: MIT, 1997).

As is well appreciated, assortment favors the intuitively “fair” or “just” strategy in a variety of different games. For example, it is one of the few plausible mechanisms that cause cooperation to evolve in the one-shot Prisoner’s Dilemma. For the cooperative strategy in the Prisoner’s Dilemma is the one that maximizes the overall payoff when it is employed by all the players. In contrast, two players who both use the uncooperative strategy will achieve only very low payoffs. Interestingly, this seems to be a general feature of the games that model prosocial behaviors. That is, the prosocial behaviors are normally the ones that maximize payoffs, under the assumption that everyone plays the same strategy.¹¹

I.3. Efficiency. A more controversial mechanism, usually considered only in a biological context, is group selection. Although there is a good deal of controversy regarding the nature of group selection mechanisms—and indeed whether there are group selection mechanisms at all—we can give them a general characterization that is adequate for our purposes here. Group selection mechanisms are said to operate when there is some environmental pressure that forces one group of conspecifics to compete with each other for survival and reproduction. As Elliott Sober and David Sloan Wilson have put it, group selection operates whenever a group of individuals are “in the same boat” as regards survival and reproduction.¹²

It is a difficult and complicated matter to model intergroup competition. But many such models have converged on the use of efficiency as a measure of what we might call “group fitness.” Efficiency, as understood by economists, is simply a measure of how little of a resource is wasted in interactions. When a group of individuals plays a set of strategies that wastes very little, then we say that the group has high efficiency. In a game-theoretic context, efficiency is measured simply by taking the sum of the payoffs in an interaction.

Indeed, this interpretation of the concept of efficiency makes sense if payoffs are interpreted as fitness. So if a group has high average payoffs, then the group will tend to reproduce more quickly. And higher rates of reproduction make it less likely that a group will face extinction, be outcompeted for scarce resources, and so on.

Of course, it is apparent that efficiency considerations will favor prosocial behaviors like that represented in the Stag Hunt game. For

¹¹ Skyrms, “Darwin Meets *The Logic of Decision*: Correlation in Evolutionary Game Theory,” *Philosophy of Science*, LXI (1994): 503–28.

¹² “Reintroducing Group Selection to the Human Behavioral Sciences,” *Behavioral and Brain Sciences*, xvii (1994): 585–654. But see Sober and Wilson, *Unto Others* (Cambridge: Harvard, 1998) for their current view, which has changed somewhat.

the highest possible efficiency is obtained to the degree to which a population tends to play the Stag strategy. Similarly, the intuitively “fair,” “just,” “cooperative,” or “altruistic” behaviors in other games such as the Prisoner’s Dilemma are also favored by efficiency considerations. We may conclude that if group selection mechanisms operate to a sufficiently high degree, and they favor efficient strategies, then we should expect the prosocial behaviors to evolve over time.

II. OPENING THE BOXES

At this point, it should be clear why game-theoretic explanations of prosocial behaviors do not open up the black boxes to examine the proximate mechanisms inside. For as we have seen regarding the Stag Hunt game, all three styles of explanation—stability, assortment, and efficiency—predict the same observed behaviors in games. Furthermore, the Stag Hunt is not an example that is contrived to demonstrate this phenomenon. Any of the standard coordination games or bargaining games that are typically the subject of game-theoretic studies will behave the same way. Because the three standard explanantia in game-theoretic analyses will tend to agree across a wide range of models, we have the possibility of deriving explanations of great generality. That is, if all of the available proximate mechanisms have the same observational consequences, then there is little pressure to examine the details of those explanations, or uncover differences between them.

There is, however, another simplifying assumption in the game-theoretic models that has gone largely unnoticed.¹³ For in each of the standard games—including the Stag Hunt, Prisoner’s Dilemma, and Nash Bargaining Game¹⁴—each player faces the same available strategies and the same range of potential payoffs. In the language of game theory, we say that such games are symmetric. The restriction to symmetric games has an excellent pedigree. In his seminal paper, Nash made this restriction so as to express the assumption that the players have equal bargaining power. However, I am not aware of any recent discussions of the evolution of prosocial behaviors that explicitly justifies such an assumption. It is this restriction to symmetric games that has obscured several important conceptual and formal questions that need to be addressed.

II.1. Ambiguity in the Explanandum. Suppose that two individuals—call them *A* and *B*—must decide on how a resource is to be divided

¹³ A notable exception may be found in Ken Binmore, *Game Theory and the Social Contract* (Cambridge: MIT, 1994).

¹⁴ Also known as “Divide the Cake” or “Divide the Dollar.”

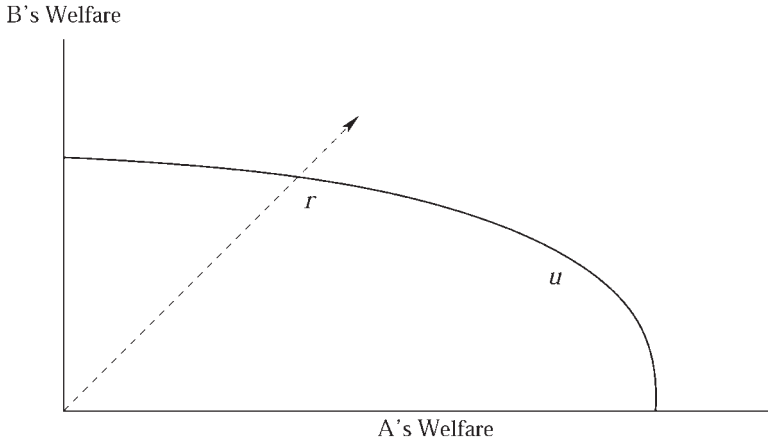


Figure 2: An asymmetric bargaining situation.

between them.¹⁵ If we relax the assumption that their choices are symmetric, then we may diagram the situation as in Figure 2. There, the possible payoffs to *A* and to *B* are represented along the *X* and *Y* axes, respectively. Points further to the right represent outcomes giving *A* higher payoffs, while points higher along the *Y* axis represent outcomes giving *B* a high payoff. The range of jointly feasible outcomes is represented by the curve; any combination that would put the payoffs of *A* and *B* under the curve are feasible bargains. As is represented in Figure 2, player *A* has the potential to enjoy higher payoffs than player *B*. We need not specify why this is so—we can say that it is due to “differences in birth and natural endowment,” as does John Rawls.¹⁶

The asymmetry of the game invites differences in opinion as to the “fair” or “just” bargain. The Rawlsian maximin solution is point *r*; for at point *r*, each player has an identical payoff, and no individual could do better without the other doing worse.¹⁷ However, a utilitarian solution will identify point *u* as the just outcome, because that point maximizes the sum of the two payoffs.

Note that if the situation were symmetric, then point *r* and point *u* would be the same. Thus, the debate between the Rawlsian and the

¹⁵ The form of this example is taken from Binmore.

¹⁶ *A Theory of Justice* (Cambridge: Harvard, 1972).

¹⁷ In general, if the range of feasible outcomes is a convex set, then the Rawlsian solution may be found simply by drawing a line at a 45-degree angle through the origin, and intersecting it with the boundary of the set of feasible outcomes (where the boundary is also known as the Pareto frontier).

utilitarian does not arise in a game-theoretic analysis unless the game is asymmetric. Thus, if symmetric games are favored in discussions of the evolution of prosocial behaviors, we should expect a large degree of ambiguity as to exactly which prosocial behavior, norm, or convention we are examining.

Indeed, it is not hard to find such ambiguity. Skyrms, for example, characterizes his project alternately as explaining the evolution of “justice,” “fairness,” and “the social contract.” He also describes the explanandum as a sense of fairness and as a moral norm, although it is unclear that a moral “sense” can be equated with a norm.¹⁸ The Prisoner’s Dilemma—which may be the most influential game-theoretic model of prosocial behavior—has been understood alternately as a model of cooperation, fairness, and most recently as a model of altruism.¹⁹ However, it should be obvious that fairness, cooperation, and altruism are not the same concepts. For instance, we may cooperate so that our individual self-interests are advanced, and this is not an altruistic behavior. Similarly, two uncooperative individuals may have a fair settlement imposed upon them. And although it is a case of cooperation when everyone agrees to drive on the right side of the road, there is nothing altruistic about that coordinated behavior.

The fact that these different concepts are frequently run together in game-theoretic analyses can be blamed largely upon the fact that the games in such discussions are typically symmetric. Already, if we use a formal tool like game theory to examine prosocial behaviors, we are necessarily subjecting the analysis to a good deal of simplification. But if we make the additional assumption that the players in such games are faced with identical choices and possible outcomes, then there is no longer any formal characteristic that can be used to pry apart alternate moral theories, or indeed, to distinguish between various moral concepts such as fairness, justice, cooperation, and altruism.

II.2. Formal Considerations. Not only is the explanandum obscured in models of symmetric games, but the explanans is also obscured to at least the same degree. To see this, we return to the example of an asymmetric bargaining situation illustrated in Figure 2. By showing how the asymmetric game allows us to distinguish among various explanantia, we may better see how symmetric games fail in this regard.

¹⁸ Although it is important to note that Skyrms does not intend for his work to constitute a full-blown theory of the evolution of justice, or any other prosocial behavior or norm. As he says in *Evolution of the Social Contract*, his models are merely intended to be “the beginning of an explanation” of those concepts.

¹⁹ See Axelrod; Binmore; and Sober and Wilson’s *Unto Others*, respectively, for examples of these interpretations.

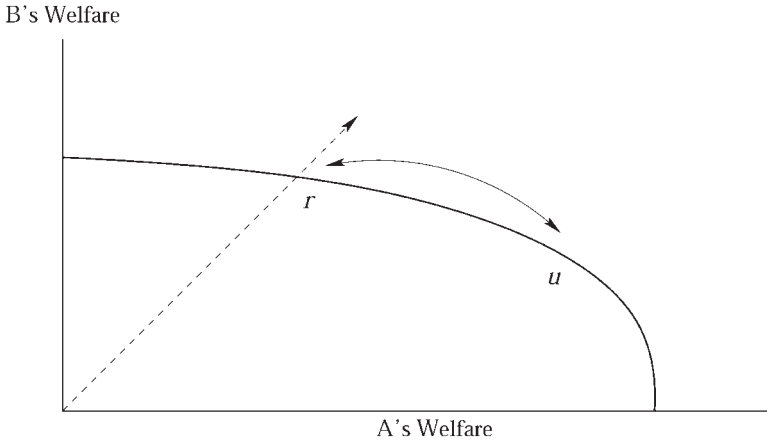


Figure 3: Group selection mechanisms should put the population somewhere between point r and point u , in the range indicated by the curved arrow.

For example, suppose that an interaction is asymmetric and occurs in the context of biological evolution and natural selection. As I have argued, if group selection mechanisms operate, then those mechanisms will tend to push the population to the efficient outcome. Because the efficient outcome in our simple example is the utilitarian solution at point u , group selection mechanisms will push the population toward that point. However, if the dominant mechanism is assortment, then the population dynamic will move the group's behavior along the dashed line toward the Rawlsian maximin solution at point r . If both mechanisms are operating to some significant degree, then it will turn out that neither point u nor point r will be stable. Instead, the stable point will be some compromise between those two points, somewhere along the curved line in Figure 3.

Even in this simple example, there are some surprising lessons that can be learned. The first concerns how we should formally model group selection mechanisms. Advocates of group selection—especially Sober and Wilson—place assortment mechanisms at the center of formal models of such processes. For group selection is said to operate to the degree to which different phenotypes become concentrated in different groups. For example, group selection will operate when there is some mechanism that tends to put altruists in one group, and selfish types in another. When such a process operates, the predominantly altruistic group will have a higher average fitness than the selfish group. Of course, if there is some form of intergroup selection operating, then the altruistic group will be favored.

In this way, assortment and group efficiency go hand-in-hand. Assortation brings about intergroup differences in group efficiency; group selection mechanisms select the more efficient group. It is uncontroversial that assortment mechanisms are correctly emphasized as a necessary prerequisite for group selection; and this is agreed upon even by those, like Richard Dawkins, who argue forcefully against group selectionist explanations.²⁰

Although it is certainly correct that assortment and efficiency do work together in this way, it is an underappreciated fact that these two mechanisms may work against each other. For as I have argued above, in an asymmetric game, assortment mechanisms will push the population toward point r , while efficiency will push the population away from point r , and closer to point u . Thus, in an environment in which group selection operates, we should not expect to find the population at either point, but somewhere in a range between those two points, as in Figure 3.

Of course, if the game is symmetric, and point u is the same as point r , then it is safe to equate assortment and efficiency, putting them both under the general heading of "group selection." But since many interactions are not symmetric, there are many cases in which assortment and efficiency should be kept apart.

Precisely where the population will end up depends upon the relative influence of assortment and efficiency. For this reason, we must pay attention to the precise nature of the mechanisms that operate. This is in stark contrast to the conventional wisdom on this topic, which asserts that it does not matter precisely which assortment mechanism is operating. Both Skyrms and Sober and Wilson cite the same passage from Hamilton, that

...it obviously makes no difference if altruists settle with altruists because they are related...or because they recognize fellow altruists as such, or settle together because of some pleiotropic effect of the gene on habitat preference.²¹

In Hamilton's example, a gene causes altruistic behavior, and causes those with the gene to settle in a particular habitat. Thus, the altruistic individuals tend to settle in one habitat, while the selfish types settle in another.

In fact, assortment mechanisms are sometimes grouped together under the heading "greenbeard effect," so called because of a fanciful

²⁰ *The Selfish Gene* (New York: Oxford, 1976).

²¹ Sober and Wilson, *Unto Others*, p. 134.

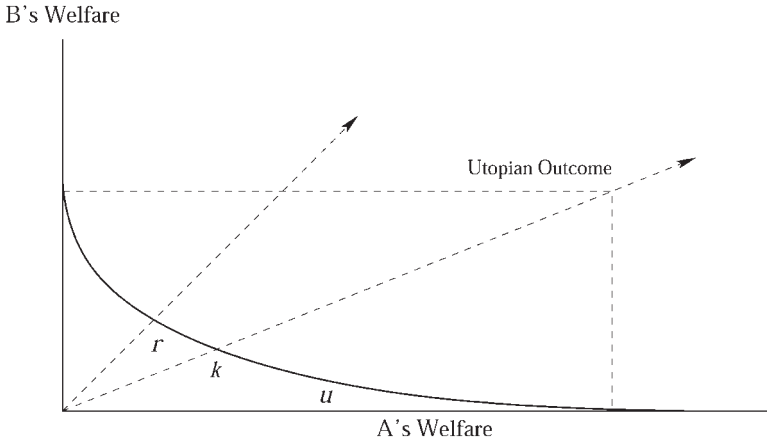


Figure 4: In a game without a convex payoff structure, the Kalai-Smorodinski, utilitarian, and Rawlsian outcomes may be distinct.

example due to Dawkins. In his example, altruists have green beards, and people with green beards prefer to interact with each other (*op. cit.*, p. 89). The point of this imaginary example of an assortment mechanism is that it does not matter what mechanism sorts the population into altruist and nonaltruist groups, so long as something does.

The considerations that I have raised above suggest, however, that we should be hesitant to treat all mechanisms of assortment the same. If different mechanisms of assortment operate to different degrees, then the population will be pushed further toward point r (if assortment is strong relative to efficiency), or closer to point u .²²

II.3. Ambiguity among Proximate Mechanisms. The points u and r corre-

²² The conflict between Rawlsian and utilitarian solution concepts in a biological setting corresponds closely to the conflict between the evolution of female-biased sex ratios and the evolution of sex ratios approaching unity. For in a group-selection model, within-group selection will favor the evolution of a 1:1 sex ratio (as in R. A. Fisher's original argument), while between-group selection will favor a female-biased sex ratio. In a group-selection model, the two pressures are in conflict with each other; accordingly, we observe that the actual sex ratios are a compromise between those two pressures. Specifically, the stronger the between-group selection, the more heavily biased toward a female sex ratio. A female-biased sex ratio produces offspring more efficiently; thus, the tendency toward a female-biased sex ratio corresponds to the efficient utilitarian solution. On the other hand, a 1:1 sex ratio corresponds to (what I have called) the Rawlsian outcome. Accordingly, the mathematics of a group-selectionist model of sex ratio evolution is analogous to the mathematics of asymmetric games under the dynamics I have informally described above. For a discussion of the evolution of female-biased sex ratios, see D.S. Wilson and R.K. Colwell, "Evolution of Sex Ratio in Structured Demes," *Evolution*, xxxv (1981): 882-97.

spond to two different moral intuitions—the utilitarian’s intuition that the fair outcome maximizes the welfare of everyone, and the Rawlsian intuition that we ought to maximize the welfare of the least well-off. Other solution concepts may be illustrated by other asymmetric payoff curves. The payoff curve in Figure 2 is said to be “convex,” because a line connecting any points inside the curve does not go outside the curve. If this restriction is lifted, then we have the possibility of considering games whose payoff structure like that depicted in Figure 4. Clearly, the utilitarian and Rawlsian solutions may be arbitrarily far apart in a game with a concave payoff structure. Additionally, it is easy to illustrate other conceptions of “fair” outcomes, such as the Kalai-Smorodinski solution. According to Kalai-Smorodinski, the fair outcome is calculated by first locating the Utopian solution, which is a state in which *A* and *B* both enjoy their maximum possible payoffs. Typically, that point will lie outside the range of possible outcomes, because *A* and *B* cannot simultaneously obtain their highest possible payoffs. Accordingly, the Kalai-Smorodinski solution concept tells us to draw a straight line from the status quo point (or original position) to the Utopian outcome.²³ The fair outcome is the best point on the line inside the range of feasible outcomes. This is illustrated by point *k* in Figure 4.

Clearly, each solution concept enjoys a significant degree of intuitive plausibility. Because different mechanisms—both biological and social—have the potential to push populations toward different solution concepts, it is important to clarify the explanandum when we set out to explain “the origins of prosocial behavior.” Instead of trying to understand the origins of a broad range of prosocial practices, intuitions, and sentiments, we should be setting out to explain the origins of particular moral intuitions or practices. For instance, we may try to explain the origins of broadly utilitarian sentiments, or we might try to explain why the maximin principle appeals to our sense of fair play.

In fact, asymmetric games allow us to highlight the different causal influences of various social and biological mechanisms. If a significant proportion of morally relevant interactions are asymmetric, and these causal factors trend toward different solution concepts, then we should expect that “our moral intuitions” are not a unified whole. Instead, our moral sentiments and practices would be the result of overlapping and contradictory social and biological influences.

²³ The Kalai-Smorodinski solution plays an important role in the theory advocated by Binmore.

Indeed, when experimental economists conduct studies of human beings' actual social behavior in bargaining situations, they model our behavior as governed by an overlapping set of conflicting norms. It should come as no shock to either conventional wisdom or moral theory that our moral sentiments, intuitions, and social practices are highly conflicted. Various environmental triggers activate different norms to varying degrees. These triggers have been investigated in detail by experimental economists, who have identified the roles played by institutions, the scale of the stakes involved, as well as cultural differences.²⁴ These norms may conflict with each other, and each is activated to varying degrees by different environmental cues. For example, Elizabeth Hoffman and others have argued that humans calculate the so-called "social distance" between themselves and their partners.²⁵ The smaller the social distance, the greater the degree of activation of norms favoring equal division in bargaining situations. Accordingly, when one's bargaining partner is a member of one's own social group, test subjects overwhelmingly tend toward equal division, even when there is a financial disincentive to one partner.

III. LESSONS LEARNED

If our moral intuitions are conflicted, and are triggered by a variety of different factors, then it is a mistake to take on the project of explaining "the origins of fairness" —or any other general prosocial behavior—simpliciter. For if the project is specified in such a way, then it incorrectly treats our moral intuitions as if they were a unified and consistent set. This mistake is compounded if it also commits us to treating all evolutionary (both social and biological) mechanisms as if they can all be described by the same dynamics.

It is an interesting change that the sociobiologists, including E.O. Wilson²⁶ and Michael Ruse,²⁷ did not think of their project as explaining the origins of fairness, altruism, justice, the social contract,

²⁴ See Bruno S. Frey and Iris Bohnet, "Institutions Affect Fairness: Experimental Investigations," *Behavior*, LXXV (1980): 262–300; Lisa A. Cameron, "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia," *Economic Inquiry*, XXXVII (1999): 47–59; Alvin Roth, V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir, "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review*, LXXXI (1991): 1068–95.

²⁵ Hoffman, Kevin McCabe, Leoth Shachat, and Vernon Smith, "Preferences, Property Rights and Anonymity in Bargaining Games," *Games and Economic Behavior*, VII (1994): 346–80; Hoffman, McCabe, and Smith, "Social Distance and Other-Regarding Behavior in Dictator Games," *The American Economic Review*, LXXXVI (1996): 653–60.

²⁶ *Sociobiology: The New Synthesis* (Cambridge: Harvard, 1975).

²⁷ Ruse and Wilson, "Moral Philosophy as Applied Science," in Sober, ed., *Conceptual Issues in Evolutionary Biology* (Cambridge: MIT, 1984), pp. 421–38.

or any other general moral category. Instead, their proposed explanations were restricted to particular social practices such as inheritance, racism, preferential treatment of genetic relatives, and so on. In contrast, current work on the evolution of prosocial behaviors tends to be much more general. Skyrms, J. McKenzie Alexander,²⁸ Sober, Wilson, and others tend to be concerned with general moral categories like “the social contract,” “altruism,” “justice,” and “fairness.”

The switch from specific and particularized explanations to highly general explanations coincides with the rise of game theory as the favored explanatory tool. If the arguments above are correct, then it is not a coincidence that the explanandum changed in this way when game theory became widely employed. After all, the standard games used to model prosocial behaviors are symmetric. Because all of these moral concepts and prosocial behaviors are conflated in symmetric games, the models encourage us to conflate all of these moral concepts. Similarly, the mechanisms that may lead to prosocial behaviors also push populations to the same outcomes, provided that the games are symmetric. Thus, the exclusive reliance on symmetric games encourages us to ignore the underlying mechanisms that are ultimately responsible for the evolution of prosocial behaviors.

There is a cost of ignoring the specific mechanisms that are supposed to lead to prosocial behaviors. Any such proposed explanation is vulnerable to the criticism that the devil is in the details. That is, if one can come up with a situation in which there is some force that is significantly different from the standard game-theoretic models, and which leads to a different outcome, then one has a substantive criticism. And of course, this is the approach that is standardly taken to object to game-theoretic explanations of prosocial behaviors. For instance, D’Arms and others have pointed out that there are cases in which we might expect an anti-correlation in interaction²⁹; Philip Kitcher has argued that the presence or absence of coalitions may play an important role in biological contexts³⁰; Neil Tennant has argued that it may make a difference whether reproduction is sexual or asexual in biological models³¹; and I have questioned elsewhere the assumption that biological and social evolution are analogous.³²

²⁸ “Evolutionary Explanations of Distributive Justice,” *Philosophy of Science*, LXVII (2000): 490–516.

²⁹ “Game Theoretic Explanations and the Evolution of Justice.”

³⁰ “Games Social Animals Play: Commentary on Brian Skyrms’s *Evolution of the Social Contract*,” *Philosophy and Phenomenological Research*, LIX (1999): 221–28.

³¹ “Sex and the Evolution of Fair-dealing,” *Philosophy of Science*, LXVI (1999): 391–414.

³² Zachary Ernst, “Explaining the Social Contract,” *British Journal for the Philosophy of Science*, LII (2001): 1–24.

This line of objection is basically sound. It is too much to expect a single model to be agnostic, not only about which mechanism is supposed to shoulder the explanatory burden, but also about what the explanandum is. Accordingly, it is relatively easy to identify potential causal influences that may adversely affect the explanatory significance of the game-theoretic models. But we should not take this line of objection to be more significant than it actually is. For there is no intrinsic feature of the game-theoretic models that forces us to neglect specifying the explanandum. In order to deal with these criticisms, game-theoretic models must be detailed enough to reflect the finer distinctions between various evolutionary mechanisms as well as the differences between varieties of prosocial behaviors. Accordingly, we must pay attention to asymmetric games.

ZACHARY ERNST

Florida State University