Robustness and Conceptual Analysis in Evolutionary Game Theory

Zachary Ernst†

A variety of robustness objections have been made against evolutionary game theory. One of these objections alleges that the games used in the underlying model are too arbitrary and oversimplified to generate a robust model of interesting prosocial behaviors. In this paper, I argue that the robustness objection can be met. However, in order to do so, we must attend to important conceptual issues regarding the nature of fairness, justice, and other moral concepts. Specifically, we must better understand the relationship between moral concepts and formal characterizations of games.

1. Introduction. Evolutionary game theory is primarily understood as a tool in the *descriptive* project of explaining the historical origins of certain behavioral propensities. For example, if we want to know how people have come to have the propensity to cooperate in certain one-shot Prisoner's Dilemmas, we may try to answer this question by exhibiting an evolutionary game-theoretic model in which the simulated agents evolve that behavior under certain initial conditions. If the model is based upon plausible assumptions, then it is taken as at least a partial explanation of why people exhibit that behavior in real-world circumstances.

This is understood as a purely descriptive project, and advocates of this approach—such as Brian Skyrms (1999, 2001, 1994a, 1994b, 1996), Elliott Sober and David Sloan Wilson (1992, 1994, 1998), and Jason Alexander (2000, 1999)—tend not to address normative questions or conceptual questions concerning issues of fairness, justice, and the social contract generally. Their avoidance of normative and conceptual questions has at least two motivations. First, they take the descriptive project to be a philosophically interesting one in and of itself. In particular, Skyrms has argued at length that the rational choice tradition, in which morality is supposed to derive from purely rational considerations, has failed to give

†To contact the author, please write to: Department of Philosophy, 438 General Classroom Bldg., University of Missouri-Columbia, Columbia, MO 65211; e-mail: ernstz@missouri.edu.

Philosophy of Science, 72 (December 2005) pp. 1187–1196. 0031-8248/2005/7205-0044$10.00

an explanation of our moral practices. This leaves a significant gap in the project, which is to be filled by the descriptive project he has advocated. Skyrms is not alone in this conviction; philosophically minded economists such as Ken Binmore (1994, 1989), Larry Samuelson (1994, 1997), and H. Peyton Young (1994) also take a dim view on the project of rationally justifying our moral practices. According to them, if we understand rationality as some kind of informed self-interest, then we have no reason to expect rational people to behave in a just or fair way, despite the arguments of rational choice theorists such as Jon Elster (1992, 1989), David Gauthier (1993, 1986), and John Rawls (1972). Another reason for avoiding the normative and conceptual questions is out of a fear of committing the naturalistic fallacy. For if one takes a strong position on the impossibility of inferring an ought from an is, then any descriptive project must be carefully distinguished from any normative one.

As a descriptive project, evolutionary game theoretic accounts are subject to robustness objections. That is, the evolutionary game theoretic models necessarily contain a large number of simplifications and idealizations, which are crucial for making the models mathematically tractable. Thus, it is incumbent upon the advocates of evolutionary game theory to show that these idealizations do not do not render the models so unrealistic that they are no longer applicable to the real world. The robustness of these models has been criticized by several authors, and this is probably the criticism that has attracted the most attention in the literature (D'Arms et al. 1998; D'Arms 1996; Ernst 2001; Barrett et al. 1999; Kitcher 1999). I believe that these robustness worries may be met (Ernst 2001). But I shall argue here that the strategy that we employ for answering the robustness objections bears directly on important conceptual issues regarding the nature of fairness, justice, and other prosocial behaviors. That is, if we cope with the robustness worries in the right way, then we may simultaneously sharpen our understanding of the explanandum behaviors.

2. Robustness in Evolutionary Game Theory. In this paper, we will not be concerned with the details of evolutionary game theoretic models.[1] For our purposes it is enough to think of an evolutionary game theoretic model in very general terms. Such models may be thought of as simulating the evolution of a large number of individuals who interact by playing a game, such as the Prisoner's Dilemma or Divide-the-Cake. The individuals are paired up, play the game, and are assigned a fitness that is based upon the payoff that they receive in their interactions. We may then think of

1. A good introduction to a variety of evolutionary game theoretic models may be found in Weibull 1995, and a quick introduction is in Sober 1993.

the individuals as having a number of offspring that is proportional to their fitness. The parent generation then dies, and the process is reiterated with the new generation. In this way, certain strategies may eventually come to predominate in the population, if their bearers enjoy a higher than average fitness. If the most successful strategies are the ones that are intuitively 'fair', 'just', or 'altruistic', then this is supposed to explain why those behaviors are observed in real world populations.

Robustness worries arise because we must make several different assumptions when we set up the model. We must determine whether the individuals are paired up at random, or according to some assortative mechanism based, for instance, on spatial proximity or genetic relatedness. Similarly, we must determine the functions that relate payoff and fitness; and we must also decide whether the dynamic is supposed to represent a social learning function or a biological process of natural selection. And we must also decide whether we are interested in the long term behavior of the model, or its behavior in the short run.[2]

More obviously, we must also choose which game the simulated agents are supposed to play. Following the work of Robert Axelrod (1984, 1997), the traditional choice has been to focus on the iterated Prisoner's Dilemma, because that game provides us with a worst case scenario for the evolution of cooperation. Additionally, Sober and Wilson have argued that although the Prisoner's Dilemma has traditionally been used as a model of cooperation, it also happens to be the best available model of altruism (1998). But more recently, Skyrms has proposed that we use Divide-the-Cake, which was introduced by John Nash (1950) as a model of strategic bargaining. Skyrms has also offered a provocative argument that the Stag Hunt game, which differs significantly from both the Prisoner's Dilemma and Divide-the-Cake, should be the focus of our investigations into the evolution of the social contract (2001, 2004). A variety of other games are commonly discussed in the economics literature (which is the traditional home of evolutionary game theory), including the so called Dictator game, the Ultimatum game, and the Centipede game (Bethwaite and Thompkinson 1996; Camerer and Thaler 1995; Gale et al. 1993; Güth et al. 1982; Hoffman 1996; McKelvey 1992). Each one has at least some claim to be a model of fairness, altruism, cooperation, or other prosocial behavior.

The existence of such a wide variety of games has been taken to comprise a robustness objection to the evolutionary game theoretic models. Clearly, if there are several plausible games that may each serve as a

2. Larry Samuelson has distinguished between what he calls the 'short run', the 'long run', and the 'very long run' equilibria in evolutionary games (1997), and has argued that each may be appropriate for different phenomena.

model of (say) fairness, and these games behave differently in the evolutionary models, then we will have to justify whichever game we model. For example, the Prisoner's Dilemma is notoriously characterized by the fact that the intuitively fair strategy of cooperation will never be the rational choice in any one shot game, for defection is guaranteed to have a higher payoff, regardless of what strategy is played by one's partner. But in Divide-the-Cake (which is the focus of Skyrms 1996), the intuitively 'fair' strategy *will* sometimes be the choice of a rational agent, for it yields the highest payoff when paired up against itself. This single difference is enough to force radically different behavior in evolutionary models of the two games. For instance, in the standard replicator dynamics model of the Prisoner's Dilemma, cooperation will invariably die out if there are any defectors in the initial population. But if we model Divide-the-Cake under identical dynamics, the fair strategy is likely to predominate in the population. And yet, if each has equal claim as a model of fairness, then one may object that an evolutionary model of either fails the test of robustness. Indeed, this objection has been raised by Philip Kitcher (1999).

The favored response to this particular robustness worry has been to posit some mechanism that favors the evolution of prosocial behavior in virtually any game. Specifically, the centerpiece of most evolutionary explanations is a mechanism of assortment, by which a player is more likely to interact with another player who uses the same strategy. It is hard to overstate how important assortative mechanisms have been in game theoretic explanations of prosocial behaviors, both in philosophical, economic, and biological work. They are the only plausible mechanisms capable of sustaining cooperation in the one shot Prisoner's Dilemma, and they have an extremely strong positive influence on other games. However, there are cases—particularly in social settings—where it is unrealistic to expect correlation. As D'Arms et al. (1998) have argued, if we are dealing with cognitively sophisticated agents, then we would not expect the most 'greedy' or 'selfish' individuals to seek each other out. Rather, we should expect them to try to establish a strong anticorrelation in the population.

Other mechanisms that favor the evolution of prosocial behaviors, while provocative, do not fare as well as assortment. If the population has some sort of spatial structure, and individuals tend to interact with their neighbors, then the prosocial behaviors may be favored (Skyrms and Alexander 1999; Young 1998). However, this depends crucially upon which game is played. In Divide-the-Cake, incorporating spatial proximity into the model has a dramatic and positive effect on the evolution of fairness. But spatial proximity is largely irrelevant to the evolution of cooperation in the one shot Prisoner's Dilemma. In other models, particularly in the context of biological evolution, we imagine a metapopulation structure

that is divided into relatively isolated subpopulations (Wright 1943, 1945). If we allow for the right amount of migration between subpopulations, then the prosocial behavior may be favored in some games like Divide-the-Cake or the Stag Hunt. But this population structure actually has the *opposite* effect in the one shot Prisoner's Dilemma—if cooperation happens to evolve in any subpopulation, it is quickly wiped out by even a single immigrant defector.

3. The Conceptual Question: Game Selection. So on the face of it, it is difficult to see how a single explanatory strategy can account for the wide range of observed prosocial behaviors in such a variety of contexts.[3] In the course of gauging how successful an explanatory strategy has been, we must determine to which class of games it applies. For example, if an explanation applies only to Divide-the-Cake, but not to the Prisoner's Dilemma, then we must be careful to distinguish between situations in which one or the other game is the most reasonable choice. Of course, the ideal situation would be to have a mechanism that works in favor of the prosocial strategy in *any* game.

At this point, we are led immediately to a prior conceptual question that has not been adequately addressed in this literature. Some games, like the Prisoner's Dilemma, are naturally taken to model prosocial behaviors. Other games are not. While many games are of theoretical interest, most are not easily interpreted as representing 'fairness', 'cooperation', or any related normative concept. The question is this: why are some games reasonable models of prosocial behaviors, while other games are not?

Another way to put the same question is to ask how we can determine, for any given game, what prosocial behavior (if any) the game models. To see that this conceptual question has not been adequately addressed, we may examine what social behaviors the standard evolutionary game theoretic models are taken to explain. Skyrms, for instance, characterizes Divide-the-Cake as a model of 'fairness', 'justice', 'cooperation', 'modesty', *and* 'the existing implicit social contract'. Sober and Wilson point out that the Prisoner's Dilemma is frequently characterized both as a model of 'cooperation' and 'altruism' (1998, 84–85). The famous strategy of Tit-for-Tat in the iterated Prisoner's Dilemma is usually called a 'nice' or 'altruistic' strategy; but someone using Tit-for-Tat will not hesitate to defect repeatedly against a defector—so this is hard to characterize as an *altruistic* strategy. Furthermore, it should be clear that the 'social contract' is something much more complex than mere cooperation, while cooper-

3. Thus, we have a *prima facie* argument against what Alexander (2000) has called 'explanatory generalism' in evolutionary game theory.

ation is not the same behavior as justice—after all, two people can cooperate in an unjust behavior. Furthermore, cooperation is clearly not the same behavior as altruism. For example, we may plausibly cooperate by agreeing to drive on the right side of the road, but this kind of cooperation has nothing to do with altruism, either in the psychological or biological sense.

Thus, there is a considerable level of conceptual unclarity with respect to the normative concepts that are being explained in the evolutionary game theory paradigm. Furthermore, as the robustness worry shows us, we cannot sweep this conceptual issue under the rug and still generate robust evolutionary explanations. For so long as the different games have different behaviors in the evolutionary models, we will always be challenged to justify our choice of game. And it is clear that this challenge cannot be met if we are unclear as to what normative concept the game is supposed to model in the first place.

The uncharitable way to interpret this gap in discussions of evolutionary game theoretic models would be to conclude that this explanatory strategy is fundamentally flawed. But I favor a charitable interpretation. Instead of taking a dim view, we could instead note that game theorists do tend to pick out the same games from a wide array of possibilities, even if they assign conflicting labels to the corresponding normative concepts. Thus, we might conclude that, although there is some level of conceptual unclarity, we are nonetheless working with some determinate normative concept. For if game theorists were really as unclear as some of their language suggests, we might reasonably expect to find a wider range of games in the evolutionary models; after all, there are infinitely many games to choose from, and we find only a relatively small number actually being discussed.

Let us call the class of games that are reasonable models of normative concepts 'fairness games'. The conceptual question is to give a reasonable set of criteria for determining when a game is in the class of fairness games, and when it is not. If we can answer that question, then we will have simultaneously made progress on two fronts. First, we will have clarified the explanandum, and thereby made progress on conceptual questions regarding the nature of prosocial behaviors like fairness, justice, and altruism. Second, we will be able to better define a strategy for dealing with robustness worries. For if we can delimit the set of fairness games, then we will have a better understanding of what conditions must be met for an explanation to be robust.

4. Toward a Characterization of Fairness Games. In what follows, I do not attempt to give a definitive characterization of fairness games, but to outline some of the considerations that must go into any such charac-

terization. Furthermore, I shall use the word 'fair' as a generic term representing the cluster of related normative concepts that are the subjects of game theoretic models.

It should be a tautology that fairness games admit of at least two distinct outcomes, one of which will be fair, and the other of which will be unfair. If we take seriously the concerns of Skyrms, Binmore, and others who criticize the rational choice tradition of reducing morality to rationality, then we must hold that it is possible that rational players could settle on an unfair outcome. For if it is impossible to give a rational justification for, e.g., cooperation in the one shot Prisoner's Dilemma, that is tantamount to saying that rationality might endorse defection. Similar considerations will apply to any fairness game—if rationality does not suffice to justify the fair choice, then rationality might endorse the unfair one. And this just means that rational players may converge on the unfair outcome.

Furthermore, we must follow the lead of economists who define rationality as a tendency to choose the highest available payoff. For any other reasonable characterization of rationality defines away the descriptive problem, by letting rational players be influenced by moral considerations or other factors that are not represented in the game's payoff matrix. So we are led to conclude that in a fairness game, a set of greedy maximizers might converge on an unfair outcome.

There are several competing theories that each characterize the set of strategies that a rational player may select. These include rationalizability (Brandenburger and Dekel 1987), Nash equilibrium (Nash 1950), and others. For simple games, these concepts will coincide. So let us take a simple definition as an example. Recall that a strategy is *dominated* just in case there is another strategy that yields a higher payoff, at least some of the time, but never yields a lower payoff. Simple arguments conclude that rational players will not select dominated strategies; and if it is common knowledge that all players are rational, then no player will play a strategy that is justified only on the assumption that someone else might play a dominated strategy. Thus, we may stipulate that a rational strategy survives the iteration elimination of dominated strategies.

So when we conclude that in a fairness game, the unfair outcome may be selected by a rational player, we are saying that:

1. A fairness game will have an unfair outcome *g* (for *greedy*) that survives the iterated deletion of dominated strategies.

Clearly, the unfair outcome cannot be the same as the fair outcome. And if rationality might not endorse the fair outcome, then it could be eliminated by deletion of dominated strategies (although it might not be). So we have:

2. A fairness game will have a fair outcome $f \neq g$, which may or may not survive the iteration of dominated strategies.

The concept of efficiency often plays a role in discussions of prosocial behaviors in games (see, e.g., Skyrms 1994a). A outcome is said to be efficient just in case it leads to the highest total payoff to all players. So in the Prisoner's Dilemma, mutual cooperation is efficient, while mutual defection is not. In fact, the formal definition of the Prisoner's Dilemma guarantees that mutual cooperation is the only efficient outcome. Intuitively, it seems that efficiency is a mark of the fair outcome. However, in other games, efficiency does not correlate with fairness. For example, in bargaining games, an outcome is efficient just in case all of the available goods are divided between the players. But the outcome in which one player gets everything is certainly not fair, even though it is efficient. A better candidate is Pareto efficiency, where an outcome is Pareto efficient just in case no player can secure a higher payoff without causing another player to receive a lower payoff. Clearly, in a bargaining game, the unique Pareto efficient outcome is one in which the divisible good is evenly split between the two players; and this coincides with our pre-analytic notion of fairness. So let us suppose that:

3. In a fairness game, the 'fair' outcome f will be Pareto efficient.

Undoubtedly, there is more to the story. But even this simple characterization gets quite close, for it corresponds quite closely with the games that are commonly studied. The Prisoner's Dilemma, Divide-the-Cake, the Stag Hunt, as well as the Ultimatum and Dictator games all qualify under this definition as fairness games. Other games, such as Matching Pennies, do not. So this conception of a fairness game is at least a reasonable place to begin.

5. Payoff. The payoff of having a characterization of fairness games is significant. First of all, it saves us from having to adopt an uncharitable interpretation of evolutionary game theoretic models, which may use different normative labels for the same behavior, or use the same game for modeling different normative concepts. We can say, instead, that there is a cluster of related normative concepts that are not equivalent, but center around a small number of shared characteristics. Our simple definition of a fairness game provides us with at least a start at characterizing what is shared by those normative concepts. Specifically, normative concepts are modeled by strategic situations in which a Pareto efficient outcome might not be selected by a rational agent, because of the existence of an alternate outcome yielding a higher payoff, and which survives the iterated deletion of dominated strategies.

This also gives us a principled way of answering robustness worries. For if a mechanism or population structure can be shown to increase the probability of having a Pareto efficient 'fair' outcome in the entire class of fairness games, then we will have addressed an important objection. Assortative mechanisms may be one such mechanism (as Skyrms (1994b) suggests), but there may be others. In a biological context, we should be looking for alternate population structures that favor the fair outcomes in fairness games. There is every reason to expect that there is a wide variety of explanatory strategies that will yield interesting results for the evolutionary game theoretic project. Clarity with respect to the conceptual issues is a prerequisite for finding these explanations.

REFERENCES

Alexander, J. McKenzie (2000), "Evolutionary Explanations of Distributive Justice", *Philosophy of Science* 67: 490–516.

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

——— (1997), "The Evolution of Strategies in the Iterated Prisoner's Dilemma", in Cristina Bicchieri, Richard Jeffrey, and Brian Skyrms (eds.), *The Dynamics of Norms*. New York: Cambridge University Press, 199–220.

Barrett, Martin, Ellery Eells, Branden Fitelson, and Elliott Sober (1999), "Models and Reality—a Review of Brian Skyrms's Evolution of the Social Contract", *Philosophy and Phenomenological Research* 59: 237–242.

Bethwaite, Judy, and Paul Tompkinson (1996), "The Ultimatum Game and Non-selfish Utility Functions", *Journal of Economic Psychology* 17: 259–271.

Binmore, Ken (1989), "Social Contract I: Harsanyi and Rawls", *Economic Journal* 99: 84–102.

——— (1994), *Game Theory and the Social Contract*. Cambridge, MA: MIT Press.

Binmore, Ken, and Larry Samuelson (1994), "An Economist's Perspective on the Evolution of Norms", *Journal of Institutional and Theoretical Economics* 150: 45–63.

Brandenburger, Adam, and Eddie Dekel (1987), "Rationalizability and Correlated Equilibria", *Econometrica* 55: 1391–1402.

Camerer, Colin, and Richard H. Thaler (1995), "Anomolies: Ultimatums, Dictators, and Manners", *Journal of Economic Perspectives* 9: 209–219.

D'Arms, Justin (1996), "Sex, Fairness, and the Theory of Games", *Journal of Philosophy* 96: 615–627.

D'Arms, Justin, Robert Batterman, and Krzyzstof Górny (1998), "Game Theoretic Explanations and the Evolution of Justice", *Philosophy of Science* 65: 76–102.

Elster, Jon (1989), *The Cement of Society: A Study of Social Order*. Cambridge: Cambridge University Press.

——— (1992), *Local Justice*. New York: Russell Sage Foundation.

Ernst, Zachary (2001), "Explaining the Social Contract", *British Journal for the Philosophy of Science* 52: 1–24.

Gale, John, Ken Binmore, and Larry Samuelson (1993), *Learning to Be Imperfect: The Ultimatum Game*. Madison: Social Systems Research Institute, University of Wisconsin.

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

Gauthier, David, and Robert Sugden, eds. (1993), *Rationality, Justice, and the Social Contract: Themes from Morals by Agreement*. Ann Arbor: University of Michigan Press.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze (1982), "An Experimental Analysis of Ultimatum Bargaining", *Journal of Economic Behavior and Organization* 3: 367–388.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith (1996), "Social Distance and Other-Regarding Behavior in Dictator Games", *American Economic Review* 86: 653–660.

Kitcher, Philip (1999), "Games Social Animals Play: Commentary on Brian Skyrms's *Evolution of the Social Contract*", *Philosophy and Phenomenological Research*, 59: 221–228.
McKelvey, R., and T. Palfrey (1992), "An Experimental Study of the Centipede Game", *Econometrica* 60: 803–836.
Nash, John (1950), "The Bargaining Problem", *Econometrica* 18: 155–162.
Rawls, John (1972), *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Samuelson, Larry (1997), *Evolutionary Games and Equilibrium Selection*. Cambridge, MA: MIT Press.
Skyrms, Bryan (1994a), "Darwin Meets the Logic of Decision: Correlation in Evolutionary Game Theory", *Philosophy of Science* 61: 503–528.
——— (1994b), "Sex and Justice", *Journal of Philosophy* 91: 305–320.
——— (1996), *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
——— (2001), "The Stag Hunt", *Proceedings and Addresses of the American Philosophical Association* 75: 31–41.
——— (2004), *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.
Skyrms, Brian, and Jason Alexander (1999), "Bargaining with Neighbors: Is Justice Contagious?", *Journal of Philosophy* 96: 588–598.
Sober, Elliott (1992), "The Evolution of Altruism: Correlation, Cost, and Benefit", *Biology and Philosophy* 7: 177–187.
——— (1993), *The Philosophy of Biology*. Boulder, CO: Westview.
——— (1994), "The Primacy of Truth-Telling and the Evolution of Lying", in Sober, Elliott (ed.), *From a Biological Point of View: Essays in Evolutionary Philosophy*. Cambridge: Cambridge University Press, 71–92.
Sober, Elliott, and David Sloan Wilson (1998), *Unto Others*. Cambridge, MA: Harvard University Press.
Weibull, Jörgen W. (1995), *Evolutionary Game Theory*. Cambridge, MA: MIT Press.
Wright, Sewall (1943), "Isolation by Distance", *Genetics* 28: 114–138.
——— (1945), "Tempo and Mode in Evolution: A Critical Review", *Ecology* 26: 415–419.
Young, H. Peyton (1994), *Equity: In Theory and Practice*. Princeton, NJ: Princeton University Press.
——— (1998), *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, NJ: Princeton University Press.