Game Theoretic Explanations and the Evolution of Justice

Justin D'Arms; Robert Batterman; Krzyzstof Gorny

Philosophy of Science, Vol. 65, No. 1 (Mar., 1998), 76-102.

Stable URL:
http://links.jstor.org/sici?sici=0031-8248%28199803%2965%3A1%3C76%3AGTEATE%3E2.0.CO%3B2-Q

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Philosophy of Science is published by The University of Chicago Press. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ucpress.html.

Philosophy of Science
©1998 Philosophy of Science Association

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2003 JSTOR

http://www.jstor.org/
Wed Sep 17 09:26:08 2003

Game Theoretic Explanations and the Evolution of Justice*

Justin D'Arms†‡
Department of Philosophy, Ohio State University

Robert Batterman
Department of Philosophy, Ohio State University

Krzyzstof Górny
Department of Physics, Ohio State University

Game theoretic explanations of the evolution of human behavior have become increasingly widespread. At their best, they allow us to abstract from misleading particulars in order to better recognize and appreciate broad patterns in the phenomena of human social life. We discuss this explanatory strategy, contrasting it with the *particularist* methodology of contemporary evolutionary psychology. We introduce some guidelines for the assessment of evolutionary game theoretic explanations of human behavior: such explanations should be *representative*, *robust*, and *flexible*. Distinguishing these features sharply can help to clarify the import and accuracy of game theorists' claims about the robustness and stability of their explanatory schemes. Our central example is the work of Brian Skyrms, who offers a game theoretic account of the evolution of our sense of justice. Modeling the same Nash game as Skyrms, we show that, while Skyrms' account is robust with respect to certain kinds of variation, it fares less well in other respects.

*Received February 1997; revised June 1997.

†Send reprint requests to Justin D'Arms, Department of Philosophy, Ohio State University, 350 University Hall, 230 North Oval Mall, Columbus, OH 43210. Or e-mail: darms.1@osu.edu.

‡We would like to thank John Doris, William Harms, Brian Skyrms, Neil Tennant, Mark Wilson, an audience at University of Pittsburgh Center for History and Philosophy of Science, and seminars at Bowling Green State University and Ohio State University for helpful discussions and criticism. We would also like to thank Elliott Sober and an anonymous referee for Philosophy of Science for detailed and helpful written comments. Robert Batterman's work has been partially supported by the National Science Foundation under Award No. SBR-952052.

Philosophy of Science, 65 (March 1998) pp. 76–102. 0031-8248/98/6501-0004$2.00
Copyright 1998 by the Philosophy of Science Association. All rights reserved.

1. Introduction. In an interesting new book and some recent articles, Brian Skyrms has proposed a game theoretic account of how norms of fair dealing, or justice, might have evolved (1994, 1996a, 1996b). In one important respect, Skyrms's work adopts the explanatory methodology of earlier work by Robert Axelrod (1984). Each of these authors seeks to give an account of the evolution of certain broadly moral aspects of human behavior at a very high level of abstraction. Each claims a kind of robustness for the strategies he explores. And each holds, in effect, that a mathematical model of the evolution of human behavior can explain our moral behavior and thought while remaining entirely agnostic about the psychological mechanisms underlying them, or the evolutionary histories from which they emerge.

Given the way this program contravenes the dominant methodological commitments of contemporary evolutionary psychologists as well as those of philosophical critics of sociobiology, it is somewhat surprising that there has been so little critical reaction. One fears that were the conclusions less anodyne, the methodology would be scrutinized more closely. In this paper, we focus primarily on Skyrms's work, though some of our claims apply to Axelrod as well. We begin by developing one of Skyrms's central examples in some detail. We show how strongly his results depend upon a form of correlation that we later call into question. In Section 3, we situate Skyrms's and Axelrod's work within a broad taxonomy of strategies of evolutionary explanation, surveying some familiar worries about evolutionary explanations of human behavior. We also articulate some general criteria which a game theoretic approach to these issues must satisfy, in order to meet its explanatory debts. In Section 4, we discuss some results that undermine Skyrms's account of the evolution of justice, using a model which, we argue, is more realistic than Skyrms's.

2. Skyrms's Account. Skyrms discusses a number of different, broadly "cooperative" strategies, in a number of different games. But his most developed model focuses on a bargaining game originally discussed by John Nash (1950). In this game, two players are to divide a cake. In the basic version of the game, each must decide independently how much of the cake to demand, without negotiation (here "bargaining game" is a misnomer), and knowing nothing about the other player. The players submit these "bids" to a neutral party, the "referee," who adds up the two bids. If, between them, the players have demanded more than 100% of the cake, neither player gets anything, and the referee eats the cake. If the bids sum to 100% of the cake or less, each player gets what she demanded.

The intuitively sensible strategy, Skyrms says (and we agree), is to

demand 1/2 the cake. This is also the strategy most people use, when the game is played in the laboratory. The sociobiological approach is to take intuitive plausibilities and widespread propensities as themselves data to be explained by evolution. Skyrms asks why this strategy seems so intuitive, and he thinks the solution lies in an evolutionary model of a competition between different possible strategies.

Imagine a population in which individuals pursue different strategies (i.e. claim different amounts of cake) in the bargaining game. For simplicity, consider Skyrms's model of a population with only three strategies: "Demand 1/2," "Demand 1/3," and "Demand 2/3."[1] Each round, individuals pair off at random and play the game. Rounds are generations, and the amount of cake an individual receives determines the number of offspring that individual sends into the next round. For simplicity, suppose reproduction is asexual. Offspring always adopt their parent's strategy. The result is that the strategies that get higher average payoffs increase their proportional representation in the population at large. If individuals demanding one half of the cake receive more cake on average than those playing any other strategy in a given round, while those demanding two thirds receive less on average, then the proportion of individuals demanding half will increase in the next round, while the proportion demanding two thirds will decrease.

Whether this happens depends upon the proportions of the population playing each strategy. 1/2ers get half the cake when they interact with each other, and when they interact with 1/3ers. 2/3ers only get paid when they interact with 1/3ers; if they meet a 1/2er or another 2/3er, they get nothing. Thus the prospects for each of these strategies are highly sensitive to who else is out there. Skyrms demonstrates that Demand 1/2 is the only pure strategy that is an ESS.[2] If any pure strategy takes over an initially mixed population, it will be Demand 1/2.[3]

1. Skyrms has also modeled populations with a greater variety of possible strategies. He says that decreasing the incremental difference (decreasing the "granularity" in his vocabulary) between strategies (e.g., allowing for strategies selecting any number of tenths of the cake between one and nine) increases the success of demand half in an appropriate sense.

2. A "pure" strategy is a strategy that always demands the same amount of cake. An ESS, or evolutionarily stable strategy, is an "uninvadeable" strategy. That is to say, if almost everyone in a population is playing it, any mutant strategy will do worse, and hence cannot invade the population through a process of natural selection. The idea derives originally from Maynard Smith and Price 1973. It is further refined in Maynard Smith 1982.

3. Clearly, whether this happens depends crucially on the initial distribution of strategies. Evolution here is driven by "frequency-dependent" selection. The fitness of a given strategy S is a function of how much cake it demands and how likely it is to get what it demands. This second factor is determined by the proportion of the population playing a strategy that allows S to get paid.

However, there is no special reason to expect that any strategy will take over the population. Another possibility is that the population will reach a polymorphic equilibrium, in which some individuals demand 1/3, and others demand 2/3. How likely are these different outcomes? This is represented graphically in Figure 1 below. Each point in the triangular space represents a possible state of the population, with different proportions of individuals playing different strategies. Vertices of the triangle represent the points at which the entire population is playing the corresponding strategy. Points on the interior of the triangle are mixed states of the population, where the relative distances to each vertex determine (inversely) the proportional representation of the corresponding strategies. The arrows indicate the direction in which the population is evolving over time. These diagrams were generated using a model we developed to test some of Skyrms's claims under a wider range of parameters. This model differs from Skyrms's in some crucial respects.[4]

It is worth describing briefly some of the significant differences between the two models. Skyrms models the evolutionary trajectories of populations from different starting points by solving the dynamical equation of the replicator dynamics. This assumes that populations are infinite, and renders the results deterministic. In our model, we do not solve any dynamical equation, but instead, pair individuals according the following scheme: First, an individual is chosen at random from the population. Thus, the probability of choosing a player of a given strategy is determined by the relative representation (relative frequency) of the strategy within the population. A second individual is then randomly paired with this first player in accordance with the new, updated, relative representation of the strategies. In other words, we sample without replacement. A round consists of these pairings until the population is exhausted. Before beginning the next round we renormalize the population so that the size of the population for the next round is the same as it was the previous round and so that the different strategies are present in new relative proportions determined by the payoffs received as a result of the interactions in the last round. Thus, the trajectories appearing in our triangular "population spaces" (see figure below) represent the evolutions of these proportions according to the scheme just outlined, and are not solutions to the differential equation of the replicator dynamics given different initial conditions as they are in Skyrms's model.

It is a virtue of our model that it allows us to track fluctuations—different games starting from identical distributions of strategies will

4. See the Appendix for a more detailed discussion of our model.

not generally follow identical paths through the population spaces. One would like to know the most probable trajectory from any given starting point. Averaging over many paths allows us to approximate those trajectories, and makes our results resemble more closely the smooth analytical trajectories of Skyrms's model. (Of course, the "most probable" trajectory is not necessarily the "average" trajectory. But some simple empirical investigations do suggest that the dispersion about the mean is genuinely small.) For what comes later, a more important virtue is that our model allows for the relatively simple introduction not only of positive correlation between the strategies, but also of negative, or "anti" correlation.

In view of all these differences, it is an interesting confirmation of Skyrms's initial results that our figures 1 and 2 reproduce quite closely the qualitative and quantitative features of the corresponding figures in his work (1996a, 15 and 20, respectively).

Figure 1 exhibits an unstable polymorphic equilibrium at point A, involving all three strategies. Mild perturbations disrupt this equilibrium, and drive the population into one of the large basins of attraction. The larger basin leads the population to an equilibrium at the pure strategy of demand 1/2. But the basin at the bottom leads to a polymorphic equilibrium at point B, where half the population demands 2/3 and half demands 1/3. Both these equilibria are strongly stable: minor perturbations will not lead the population out of either basin.

While the basin of attraction toward demand 1/2 is larger than the

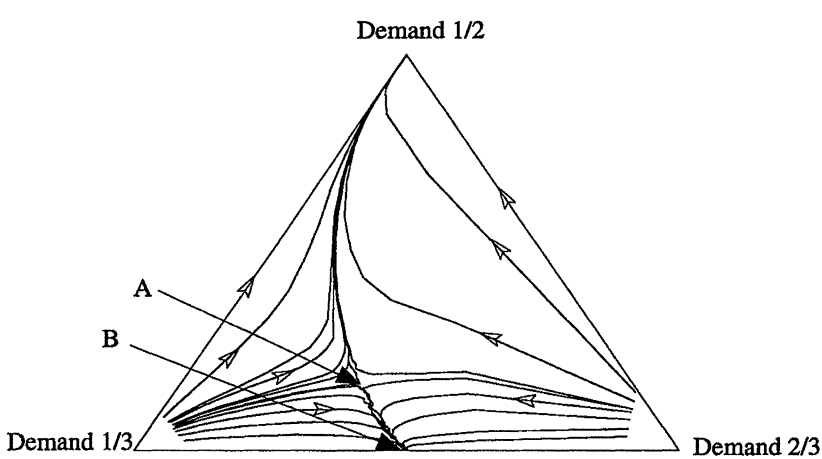

Figure 1. Correlations: 0, 0, 0

basin leading to polymorphism, the latter is substantial. Thus, if we take every possible initial state of the population to be equiprobable, we can say fair division is a more likely outcome than the polymorphism. But there is a substantial chance that a population with initial proportions selected at random will evolve toward the polymorphic state where half the population is greedy and half is modest. (This is bad news for justice, and for the population.) So far, then, the explanation of our propensity to demand 1/2 is statistical at best. In fact, it looks like the best one can hope for is an Inductive-Statistical (I-S) type explanation. That is, we explain why the population will evolve to fair division by demonstrating that such an outcome of the evolutionary process has a high probability. For Hempel one has an adequate explanation, other things being equal, when this probability can be shown to be greater than 0.5. Of course, this raises various questions about the soundness of the I-S strategy.[5] In the present context, the main issue is whether showing that the explanandum has probability greater than 0.5 would be sufficient for explanation, other things being equal. One would ideally like to show that the strategy demand 1/2 takes over the population with probability 1. Such a claim, backed by an ergodic-type limit theorem, would allow for an I-S like explanation of the success of fair division.[6] Unfortunately, it does not appear that anything like such a theorem is forthcoming.

On the other hand, Skyrms shows that there are ways to generate a much more robust conclusion. He provides graphical/numerical evidence that with the introduction of positive correlations, we may very well expect that some probability one claim is lurking in the mathematical background. Suppose that pair formation is not perfectly random. If an individual is somewhat more likely to meet another playing the same strategy than would be the case if pair formation were random, then the evolutionary trajectories will look quite different. Skyrms uses a correlation coefficient e to inflate the probabilities of like meeting like as follows:[7]

$$\mathrm{p}(S_i|S_i) = \mathrm{p}(S_i) + e\mathrm{p}(\text{not-}S_i)$$

5. See Railton 1978, 1981 for a discussion of some of the problems with the I-S model.

6. See Batterman 1992 for a discussion of this issue. It is argued there that showing high probability—namely, probability 1—is sufficient for an I-S like explanation, if the probability one claim is backed by an appropriate ergodic/limit theorem. Probability here is used in a measure-theoretic sense, so probability one does not mean certainty, and probability zero is not impossibility.

7. The introduction of correlation is somewhat more complicated in our model. See the appendix for details.

Positive correlations of this sort strongly favor demand 1/2, since that is the strategy which receives the most cake when it plays itself. The results of introducing such correlation into the model are dramatic. When *e* is greater than or equal to .2, the polymorphism virtually disappears, and every initial state of the population mixing all three strategies leads to equilibrium at demand 1/2. This is demonstrated in the figure below, where correlation coefficients are given for the strategies demand 1/3, demand 1/2, and demand 2/3, respectively.

Thus, in the correlated replicator dynamics, Skyrms's result is extremely *robust* in the following sense: it does not matter what the initial population distribution looks like—almost every population will eventually become a population of what Skyrms calls "fair dealers." It is also extremely *stable*, in that, once the equilibrium at demand 1/2 has been reached, it is highly resistant to invasion by rival strategies. Should a pocket of greedies and modests invade the population, it will quickly be eliminated. Skyrms says:

> In a finite population, in a finite time, where there is some random element in evolution, some reasonable amount of divisibility of the good and some correlation, we can say that it is likely that something close to share and share alike should evolve in dividing-the-cake situations. This is, perhaps, a beginning of an explanation of the origin of our concept of justice. (1996a, 21)

For the present discussion, we will bracket our substantial doubts about the relationship between a tendency to demand 1/2 in this bargaining game and a concept of justice.[8] Our concern here is with the nature and adequacy of the explanatory claim. How plausible is Skyrms's account as an explanation of our thought and overt behavior with respect to situations having the structure of this game? We now digress for some general observations about evolutionary explanation, which later will be relevant to evaluating Skyrms's program.

3. Evolutionary Explanations of Behavior. Any attempt to explain human behavior by appeal to evolution confronts some familiar difficulties. The best known critiques of human sociobiology, by Philip Kitcher (1985), and Steven Jay Gould and Richard Lewontin (1979), argue that sociobiological accounts often fail for abstracting from crucial details of the particular ecologies and ontogenies of the creatures whose behavior they seek to explain (see also Sober 1993, Sterelny 1992). We

8. See D'Arms 1996 for a discussion of these issues. See also Gibbard 1982 for an argument that evolutionary explanations of justice should seek to explain the moral sentiments surrounding this notion rather than substantive principles of justice.

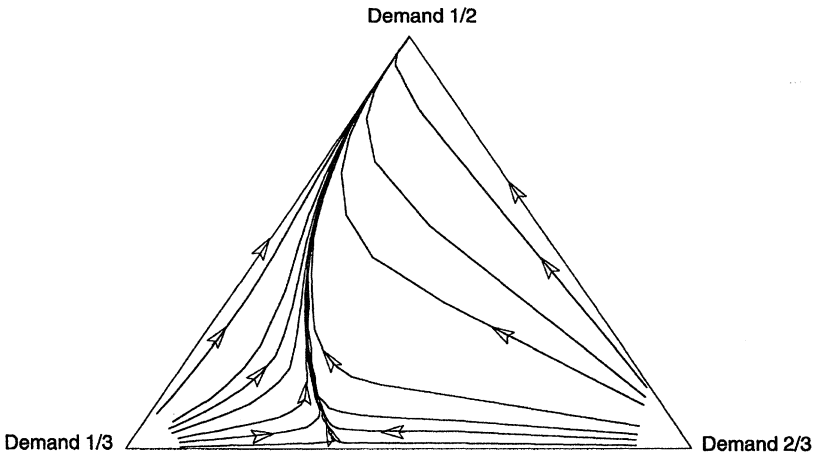

Figure 2. Correlations: .2, .2, .2

think these criticisms hit the mark against some of the work of such early sociobiologists as E. O. Wilson, Richard Alexander, and David Barash. Part of what renders these authors particularly vulnerable to such charges is that they often produce no concrete description of the mechanisms evolution has allegedly forged to produce the behaviors they claim to explain. They leave out too many of the details that would figure in an ideal explanatory text (Railton 1981). Instead, they offer accounts of how some (often apparently maladaptive) behavior might be produced by natural selection. This methodology is quite problematic when the explanations in question seem to require finely calibrated "strategic" behavior, and the considerations relevant to such a calculation are not available to consciousness. We often want to explain our behavior by appeal to the reasons for which we reach the practical conclusion that guides us. When sociobiological accounts seek to supplement or supplant rational or cultural explanation of such intentional human behaviors, and especially when such explanations undermine our own understanding of our practices, it is appropriate to request an account of how facts about fitness have impinged themselves on the agent. Failure to provide such an account is not a decisive objection to the explanation, but (we think) can often be counted against it.[9]

9. Of course, this objection to sociobiological explanation only applies to explanations of intentional behavior. There are a number of complicated issues in this area which we will not explore here.

i. Evolutionary Particularism. One response to these objections to the adaptationist program has been a move toward a *particularist* approach to evolutionary explanation of human behavior. This is the approach of evolutionary psychologists and others who seek to elucidate detailed evolutionary explanations of specific behavior types in relatively circumscribed situations.[10] According to the particularist hypothesis, the human mind comprises an array of discrete adaptive mechanisms, generated through a process of natural selection in which distinctive sorts of adaptive problems forged functionally distinct adaptive solutions. [Problems which have been explored by evolutionary psychologists and others include choice of mates and sexual partners (Symons 1979, Buss 1994), spousal violence (Daly and Wilson 1988), cheater detection (Cosmides and Tooby 1992), intergroup and intragroup relations (Cosmides and Tooby 1992), and language acquisition (Pinker and Bloom 1992, Pinker 1994).] These mechanisms are functionally specialized to process information concerning specific adaptive problems and produce behavior that solves those problems. Evolutionary psychologists frequently refer to such mechanisms as "modules." Thus, for instance, the particularist hypothesis with respect to our moral capacities holds that selective pressures deriving from the fitness consequences of various social relations such as cooperation, reciprocity, coalition building, and competition for social status, have forged similarly specific adaptive psychological mechanisms which mediate cognition and motivation in these domains.

It is not yet clear whether particularists can avoid the difficulties that confront other styles of evolutionary explanation. One problem is that evolutionary psychologists have not been very clear about what their claim of modularity amounts to. In philosophical circles, most writers have followed Fodor (1983) in treating modularity primarily as an attribute of input systems. Thus it is sometimes claimed that in order for something to count as a module, it must be informationally encapsulated, cognitively impenetrable, mandatory, or exhibit other features of Fodor modules.[11] Clearly, on that conception, even psychological mechanisms adapted to very specific tasks can fail to be modules. But

10. The best general discussion we know of the strategy of evolutionary psychology, and the most detailed justification of its methodology, is Tooby and Cosmides 1992.

11. Kim Sterelny (1995), for example, argues against Daly and Wilson (1988) that sexual jealousy is not an adaptive system, on the grounds that judgments about paternity do not seem to be informationally encapsulated or free from central control. But (granting arguendo Sterelny's claim about paternity judgments) the question of whether the jealousy syndrome as a whole is an adaptive system is surely not settled by establishing that we do not have a dedicated module for judging paternity. Even if the capacities by which we assess paternity are general devices of inference exercised on evidence (from

it seems plausible that defenders of the adapted mind have a more general conception of a "module" in mind, according to which a module is something like a subroutine: a functional system that can be plugged in or out of a larger system without interfering much with that larger system's operations.[12]

Tooby and Cosmides (1992) may have obscured matters somewhat by calling their modules "domain specific" (a term that gets a technical meaning in Fodor's vocabulary). But that term too has various senses. Although it is common to treat domain specificity as a constraint on the things a cognitive system can take as input (the visual processing system is domain specific in that it only accepts light as input), one can also regard a cognitive system's domain as being specified by the conclusions it can reach. Thus, for example, Buss (1994) posits a cognitive system that amounts to a reproductive value detector module as part of the syndrome of male sexual attraction. If he were right, men would have a device which accepts various sorts of evidence depending upon culturally contingent variables (including evidence of health, youth, fertility, social status) but issues in conclusions that are all about fertility. Whatever one thinks about Buss' hypothesis, if such a mechanism did exist it would seem appropriate to describe it as domain specific in virtue of the narrow range of conclusions it issues. It seems a natural extension of this idea to treat a larger functional system as domain specific in virtue of the kinds of *behavior* it issues. Thus, even if the belief-fixing capacities underlying human courtship behavior are not domain-specific with respect to the data they accept or the conclusions they issue, the mechanisms motivating and mediating that behavior issue in actions of particular sorts which might be sufficiently specific and identifiable to constitute a "domain."

The best particularist work derives strength from a variety of methodological features, as follows: Begin with a discrete adaptive problem of likely import for our fitness. Develop an account of the kind of

---

similarity, gestation period, cigarette butts, etc.), and even if those capacities were selected for through processes having nothing whatever to do with their effects on our skill at paternity judgments, there is no reason an adaptive motivational system cannot make use of them. Similarly, evidence that we do not have a food-detector module, and must be taught which things to eat, would not undermine the claim that hunger is an adaptive system.

12. Part of the thrust of Sterelny 1995 is a demand for a detailed account of such a notion from evolutionary psychologists. What exactly does the claim of modularity, or "special-purpose mechanisms" require? The best research strategy may be to remain agnostic about this for now. Certainly, evidence that the behavioral tendency is coded for in some set of genes, or grounded in some identifiable area of the nervous system, would help the claim. But there may be room for other notions of mechanism, too—though it is hard to see how to formulate them at this stage.

psychological mechanism which could plausibly be part of human psychology, and could be dedicated to the solution of this problem. If it is to convince reasonable skeptics, evolutionary thinking about adaptive problems facing our ancestors should allow us to generate hypotheses about adaptive mechanisms that are sometimes surprising, rather than simply offering "ultimate" explanations of claims about human nature we already believed on independent grounds.[13] Now test these hypotheses empirically to see whether the predicted "adaptive" behavior occurs in the relevant context. Cross-cultural experiments should be conducted to establish that the hypothesized mechanisms are part of a generally shared repertoire of human capacities.[14] If the hypothesis does not allow us to predict and establish new truths, it can gain plausibility by explaining discoveries that are not yet well integrated into existing scientific or common sense theories. Equally importantly, if the hypothesis is that an adaptive mechanism exists to produce behavior B in circumstances C, because B-ing in C was typically fitness-enhancing in the environment of evolutionary adaptation ("EEA"), one important test of the hypothesis will be to establish that people B in C even when doing so is not adaptive, or is not recommended by various normative theories of rational choice. Nonadaptive or non-normative instances are one crucial way to establish that the behavior is the product of some specialized adaptive mechanism, rather than of some more domain general assessment of costs and benefits.

13. Thus, for example, Daly and Wilson (1988) begin by demonstrating a tendency for step-parents to be more likely to abuse or kill step-children than are their biological parents. This by itself is perhaps not surprising, inasmuch as stories about "wicked step-parents" are a familiar part of many cultures' lore. But why is this such a recurring theme in human life? A "cultural" explanation might have it that this is due to a difference in bonding opportunities. Step-parents come into a child's life late, and may sometimes fail to develop the attachments which biological parents get in the early days of a child's life. But Daly and Wilson argue that there is some more sinister mechanism calibrating parental attachment to relatedness. This leads them to predict and then substantiate a much more surprising result: the effect remains even when comparing step-parents and biological parents who have had the same opportunities to bond with the children. Stepfathers who were present during childbirth and in the home throughout infancy are still more likely to abuse. Biological fathers who were in the military or in jail during the early months or years of the child's life, are still less likely to abuse.

14. Of course, some adaptations within our species may not be universal (because of isolation of breeding populations, for example, or because of frequency-dependent selection). Still, evidence of universality, where it is available, is one way for evolutionary psychologists to contest or supplement certain sorts of cultural explanation. When a behavior appears in a range of distinct and/or comparatively isolated cultures, an adaptive explanation gains some credence.

ii. Evolutionary Generalism. Contrast the approach of these evolutionary psychologists with that of explanatory *generalists*. Generalists seek to explain behavior by pointing to adaptive advantages for those who engage in it, without attempting to explain how exactly tendencies to behave in the relevant way are embodied in a psychology.[15] A nice example is Richard Alexander's work, borrowing from the ideas of William Hamilton on inclusive fitness. Alexander sets out to explain a wide range of social behavior by appeal to the effects of kin relationships on the genetic interests of individuals. Thus, for instance, he suggests that the phenomenon of the avunculate[16] is a consequence of a social environment in which males have comparatively low confidence of paternity—so that a maternal uncle is, on average, more closely related to a child than its mother's husband is. But he offers no account of how exactly these facts about relatedness impinge themselves on agents or societies so as to bring about the set of institutions in question.

While philosophers and other critics of evolutionary attempts to explain human behavior have had little sympathy for Alexander's program, they have tended to be much more gentle toward game theorists, population geneticists, and other generalists who employ quantitative or technical approaches to these issues. Thus, for instance, even Philip Kitcher's sweeping critique of "pop sociobiology" in *Vaulting Ambition* leaves the work of William Hamilton, John Maynard Smith, and Robert Axelrod unscathed. Kim Sterelny argues (in a review with which we are largely in sympathy) that Axelrod's work shows that generalism can sometimes be an appropriate explanatory approach.

> Both analysis and observation show that evolutionary processes are more resilient than Gould and Kitcher suppose . . . Axelrod, for example, shows that 'tit-for-tat' is *robust*. . . . Axelrod shows the merits of this practice over a considerable range of environments: it is not sensitive to small local variations.[17] (Sterelny 1992, 159)

Skyrms's approach is generalist in just the way that Axelrod's is. What the generalist approach to evolutionary explanation lacks in detail, it seeks to compensate for with robustness. Rather than offering

15. The importance of the difference between explaining behaviors and explaining mechanisms is urged in Sober 1993 (especially pp. 198–199) and Sterelny 1992.

16. The avunculate is a fairly widespread practice among certain cultures in which a child's maternal uncle provides much of the child's support. Often, this uncle provides more support than the child's mother's spouse, or putative father.

17. But note Sterelny goes on to say that " 'Tit-for-tat' is more theoretical analysis than field report. . . ." Indeed, we have doubts about Axelrod's central application of it to the field (of battle).

a detailed account of the specific psychological mechanisms underlying the behaviors they seek to explain, generalists can argue that the details are unimportant. They model various strategies under various parameters, and look for evolutionarily stable strategies, or attracting equilibria. The fact that the strategies they model can defeat all sorts of rivals under all sorts of possible initial distributions, they suggest, makes our tendency to use these strategies inevitable. A more finely grained explanation, they could add, would sometimes only obfuscate. It would suggest that the details matter, when they don't. If we didn't realize the relevant behaviors through these particular mechanisms, other mechanisms would have produced them. If genetic selection at had not filled the gap, cultural evolution could have. The particularist misses the point, by ignoring the robust stability of certain behavioral strategies.

Skyrms is particularly explicit about this idea. He suggests that the inevitable pull toward demand 1/2 in the replicator dynamics offers an explanation of our tendency to demand 1/2 which does not depend even on Darwinian evolution.

> [Demand 1/2's] strong stability properties guarantee that it is an attracting equilibrium in the replicator dynamics, but also make the details of that dynamics unimportant. Fair division will be stable in any dynamics with a tendency to increase the proportion (or probability) of strategies with greater payoffs, because any unilateral deviation from fair division results in a strictly worse payoff. For this reason, the Darwinian story can be transposed into the context of *cultural evolution*, in which imitation and learning may play an important role in the dynamics.[18] (1996a, 11)

How ought we to assess these suggestions? Given the generalist's eschewal of proximate mechanisms, how should we think about whether a game theoretic model of the evolution of some particular behavior counts as an *explanation* of that behavior? The variety in the kinds models and the way they are deployed make it unlikely that necessary and sufficient conditions for adequacy could be articulated.

18. To be sure, at the point where this claim is broached, the discussion is focused on pure strategies, and polymorphisms have not yet been introduced. Thus, Skyrms goes on to acknowledge that the introduction of polymorphisms undermines the stability of the demand half equilibrium, in one sense. What matters for his evolutionary account is that the basin of attraction toward demand half remains high in realistic models. Thus, it is important that the explanation be "flexible" enough to embrace processes of cultural as well as genetical evolution. (Of course, no one these days thinks that cultural and genetical evolution are completely distinct processes, and Skyrms needn't claim anything like that.)

However, we suggest the following as initial guidelines for assessment of evolutionary game theoretic explanations of human behavior: such explanations should be representative, robust, and flexible. We discuss each of these briefly, in turn.

a. Representativeness. Circumstances with the structure of the mathematically characterized interaction which the model treats must be realized with sufficient frequency in the EEA.

Game theorists often devote rather less attention to demonstrating that their games accurately model actual human interactions than one could wish. When they attempt this task at all, their claims about the relevant payoff structures are often based on hasty assumptions. Axelrod, at least, attempts to demonstrate representativeness in one central example. He argues at some length that trench warfare exhibits the structure of a prisoner's dilemma. While we remain unconvinced by his argument, Axelrod is to be credited for recognizing the importance of the representativeness claim for his conclusions about the explanatory role of tit-for-tat.[19] Only if circumstances with the structure of a prisoner's dilemma have been a frequent part of human life can the success of tit-for-tat in computer tournaments be offered as an explanation of human behavior in situations of that structure.

For better or worse, the prisoner's dilemma has been widely accepted among philosophers as teaching us something important about ordinary conduct. The same cannot be said, however, for divide-the-cake. Furthermore, the latter game has at least one feature that seems to lack any natural analog: the referee. Accordingly, the representativeness issue seems quite pressing for Skyrms's account.

Skyrms says very little about what aspects of ordinary life have the shape of divide-the-cake. Circumstances in which we actually divide a windfall by this procedure, with a referee standing by, are pretty rare these days, and we see little reason to suppose they were more common during the Pleistocene era. On an inclusive reading, though, one might suppose that the game models many cooperative situations. If we have cooperated to secure some divisible good, then we must find a way to divide it. What then of the role of the referee? Failure to submit bids that sum to 100% or less could be taken as failure to reach agreement. One possibility is that the good spoils while we stand arguing over it.

19. See Axelrod 1984, Ch. 4, especially p. 75. During trench warfare in World War I, soldiers on opposing sides often refrained from shooting to kill, except when retaliating for casualties inflicted by the other side. But, despite Axelrod's contentions, it is not clear that this "live-and-let-live" system resembles tit-for-tat in a prisoner's dilemma, because, from the point of view of the men on the front, there is no obvious cheater's payoff: no incentive to defect by shooting first.

Or perhaps failure to agree issues in a costly fight—costly enough that, whatever the outcome, the value of the good to be divided is negligible in comparison. Indeed, the basic structure of the problem can arise even before we have secured the good: we have an opportunity, but we must agree on how to divide the profits before we can act collectively to seize this opportunity. On a wide reading, then, these circumstances are perhaps reasonably common, and the model may secure representativeness. Notice, though, that these scenarios typically arise (as, presumably, did most human interactions in the EEA) among individuals who are acquainted with one another. Thus the wide reading suggests relaxing the requirement that interactions are random. After all, acquaintance can yield information about the likely strategies of a potential cooperative partner, and intelligent players will use that information to select a partner with/from whom they can profit.

Notice further that the wide reading suggests that the game might be understood as a model of other animals' interactions, as well. Another place to look for evidence of demand half's success would be in the division of prey among social carnivores, where some of the conditions above are also met. If social carnivores typically do not share equally in their spoils, this may be thought to undermine Skyrms's explanation. If they do, that would offer it some support.[20]

b. Robustness. The desired result is achieved across a variety of different starting conditions and/or parameters.

Different models appeal to different sorts of robustness. Axelrod's claim to robustness rests upon tit-for-tat's success against a variety of different strategies, in various computer tournaments. Skyrms can claim several distinct sorts of robustness. Not only does demand 1/2 do well in trials with different granularity (more and less finely grained demands for cake); but also the size of the basins of attraction he demonstrates in the correlated replicator dynamics establish that demand 1/2 thrives from a host of different possible initial frequencies. The demonstration of robustness, then, is the great strength of Skyrms's model.

Unfortunately, most authors who invoke robustness as an explanatory virtue are not very clear about exactly what makes the allegedly robust feature robust. One way of thinking about different kinds of robustness claims involves appeal to an appropriate notion of stability under perturbation or variation. Thus, the feature is robust if it is re-

20. We have not been able to find any evidence on this point after an admittedly brief search of the foraging literature. Our lifelong surveys of nature documentaries, suggest, however, that division of prey is typically unequal, at least among lions.

alized from a wide variety of different "starting conditions": It is stable under perturbation of the starting conditions. Of course, to assess a claim of robustness we need to have some idea of what the appropriate starting conditions are. Different sorts of starting conditions will lead to different claims of robustness.

Consider Skyrms's claim that the result of introducing a small amount of positive correlation in the divide-the-cake game demonstrates the robustness of the strategy demand 1/2. This claim is represented by the fact that the entire population space constitutes the basin of attraction of the demand 1/2 strategy. In terms of stability this means that one can perturb the initial distribution of players in the populations quite considerably and still realize the same end evolutionary result. In this sense demand 1/2 with positive correlation for all strategies is more stable under perturbation, than it is without any correlation. Compare, again, Figure 2 with Figure 1.

Another way of understanding the significance of Skyrms's results under correlation is that fair dealing will emerge even if we alter the details of the dynamics. This is a different kind of robustness. In effect, the idea is that, by introducing correlation we can perturb the very dynamics itself (not just the initial conditions, but the equations governing the interactions), and still find the same behavior emerging from a variety of starting conditions. Technically, this is related to the topological notion of structural stability studied by mathematicians interested in dynamical systems and so-called "global analysis" (Smale 1980).

c. Flexibility. (i) The evolutionary strategy whose adaptiveness the model demonstrates is potentially realizable by a number of different mechanisms. (ii) The model itself can be understood to represent different possible processes.

By pointing to specific proximate mechanisms, particularists fill in some of the explanatory details that generalists leave out. Lacking any such account, generalists must make a virtue of necessity: their accounts seek plausibility through agnosticism about the details—they are in principle realizable through any of a multitude of possible mechanisms.

We have seen that Skyrms points to another kind of flexibility as well. He suggests that his model is agnostic as between a variety of processes by which demand 1/2 might defeat rival strategies. The evolution of genetic propensities for some psychological mechanism is just one set of possible processes. The replicator dynamics might instead be taken to model the choice of strategies by rational deliberators, who attempt to maximize their share of cake. If these deliberators knew how the different strategies had fared in previous rounds, and what

proportion of the population had been playing each strategy, Skyrms's model shows that over time more and more of them would converge on demand 1/2. Or we might take the model to be an account of how various possible norms compete for the allegiance of a society, with "fair dealing" gradually coming to win out against norms of "modesty" and "ambitiousness."

4. Problems with Skyrms's Account. Equipped with the guidelines above for the assessment of generalist attempts to explain human behavior, we can begin a critical examination of Skyrms's account. Recall the central role played by correlation in Skyrms's model of the bargaining game. Under the assumption of random variation, a sizable basin of attraction pulled the population toward the greedy—modest polymorphism. Once Skyrms added a small correlation coefficient, however, this polymorphism disappeared, and every mixed state of the population evolved toward fixation at demand 1/2.

What is the justification for adding a correlation factor, though? Once Skyrms relaxes the requirement of random interactions in the population, and allows some degree of assortative interactions, we need to hear a justification for assuming that the likely departure from random interactions will be toward correlation in particular. Why think that individuals are especially likely to meet others playing the same strategy as they play? Skyrms has rather little to say about this. He suggests that "because of the nondispersive nature of the population, like tends to mate with like."[21] This may be true where strategies are influenced by genes—biological populations typically show some measure of genetic clustering. But it is problematic for Skyrms to justify the introduction of correlation on these grounds, for two reasons. First, he has given us no reason to think that we have genetic proclivities for strategies in this bargaining game. Furthermore, for Skyrms to suggest that we do would involve an uncomfortable amalgam of generalist and particularist explanatory schemas: insisting on an innate biological disposition toward a strategy, without offering any concrete account of or evidence for the psychological mechanisms that subserve it.[22] Second, as we have seen from the earlier passage, Skyrms is committed to an explanation which need not proceed by Darwinian evolution. To

21. Skyrms 1996a, 17. This suggestion is broached in the context of a discussion of sex ratios, but Skyrms clearly intends an analogy between that context and the evolution of justice.

22. This is, in effect, the explanatory strategy of behavioral genetics. But behavioral geneticists recognize that this strategy requires them to offer evidence for their claims of heritability—and they attempt to do so (though some contest this evidence).

explain correlation entirely in terms of genetic relatedness is to abandon the explanatory flexibility of the account.

Perhaps, though, there are other plausible justifications for an assumption of correlation which would apply as well to non-Darwinian processes of evolution. Appeals to non-Darwinian, or "cultural" evolution, are ambiguous. One central issue in disambiguating them concerns the mechanisms of inheritance and proliferation. It is useful to distinguish a strict sense of "cultural evolution" from a looser sense. In the strict sense, formally developed by Boyd and Richerson (1985), patterns of behavior can evolve only if they are explicitly encoded in memory, and expressly taught to or mimicked by the next generation. Here "reproduction" must proceed by "social learning."[23] We will call this strict sense "SCE."

In the looser sense of cultural evolution,[24] ideas, norms, commitments, and the associated behaviors are said to succeed or fail in a kind of competition for the allegiance of persons. No particular licensed processes of transmission or "reproduction" are set forth; instead, these can include anything from explicit instruction to a tendency to strike rational agents as plausible, or compelling, or attractive, for whatever reason. Here "cultural evolution" is the process by which cultures come to accept new norms or ideas and reject old ones, and by which some such things become enshrined. We'll call this sense the "rational deliberator dynamics."

Because SCE requires social learning of strategies, there is some justification for assuming a degree of positive correlation under its dynamics. Parents instruct their children, youngsters mimic their elders, and individuals raised together are both more likely to interact and more likely to have learned the same strategies or norms. But Skyrms does not want to be committed to the claim that our tendency to demand 1/2 is the product of some explicit instruction that we all receive, or of direct mimicry. While strict cultural evolution is more friendly to Skyrms's correlation assumption, it is clearly the loose sense of cultural evolution that offers a more flexible and realistic explanatory scheme.

How might strategies be expected to evolve among rational deliberators, who choose strategies to play on the basis of expected payoff (where this depends crucially on what they expect others to do)? Here choice of strategy is not fixed by any particular genes a player can be

23. This is a technical term. Roughly, it is the transmission of stable behavioral dispositions by teaching or direct imitation (Boyd and Richerson 1985, Ch. 3).

24. Derived from Campbell 1975. This is the sense that's common in philosophical and historical literature, and in keeping with the discussion of memes in Richard Dawkins's work. See Dawkins 1976, 1982.

presumed to share with those around her, nor by whatever explicit instruction she has received. Is there any reason to suppose that under such conditions individuals playing the same strategy are more than randomly likely to interact with each other?

It could be claimed that anyone playing the bargaining game in a population where most others demand 1/3, is likely to demand 1/3 as well, because we mimic the behavior of those around us. But surely that's not the only possibility. The more seriously we want to take the idea that the players in these games are rational agents, the more we should look for ways of learning from the environment other than simply imitating the strategies of (the majority of?) those nearby. Faced with a population where most individuals are modest, for example, wouldn't a rational deliberator be likely to settle on the greedy Demand 2/3 option? On the other hand, in a population where most people demand 1/2, it makes best sense to do as they do. Thus, whether it is rational to mimic those around one (to "correlate") or to choose a different strategy (to "anti-correlate") depends on what others are doing.

We ran a number of simulations using our model, in which we relaxed the requirement of random interactions in various ways. Rather than treating correlations as an exogenous constraint imposed by genes or environment, we explored interpretations of the model in which correlating could itself be a rational part of an individual's strategy. First, consider positive correlation. If individuals are choosing strategies to play, and are able to influence the chance that they play a like-minded individual, should they do so? In Skyrms's model, when correlation is introduced, everyone correlates. This makes sense for those who demand 1/2, since they get paid when they meet each other, and they want to avoid encounters with 2/3ers in which they will not be paid. But 1/3ers have no self-interested reason to correlate—they are always rewarded, regardless of the strategy their partner plays. 2/3ers also have no reason to correlate—playing themselves, they get nothing. Figure 3 shows what happens when 1/2ers correlate by 0.2 (the amount of the across-the-board correlation in Figure 2), while 1/3ers and 2/3ers remain uncorrelated.[25]

25. In our model, correlation works as follows. The computer randomly chooses the first member of a pair based on the frequencies of the strategies in the population at that point in the round. Say this member plays strategy S_i. It then chooses the second member of the pair. If S_i has an associated correlation factor, this is used to augment (or decrease, in the case of anticorrelation) the likelihood that the second pair member will also play S_i. The round continues until each member of the population has been assigned a strategy and paired off. The formula for deflating the likelihood that a strategy S_i will play S_i (the formula, that is, for anticorrelation) is the following: $p(S_i|S_i) = p(S_i) + e_i p(S_i)$ where $e_i < 0$. Thus, for perfect anticorrelation ($e_i = -1$) the prob-

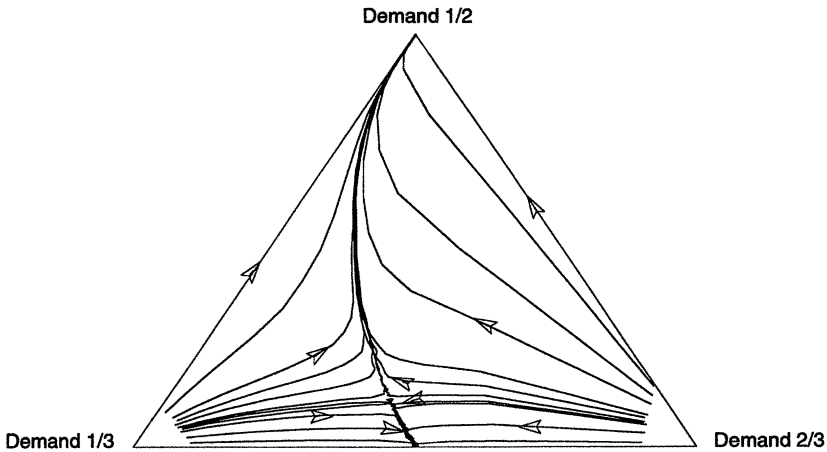

Figure 3. Correlations: 0, .2, 0

The polymorphism has reappeared! When 2/3ers are not required to pursue interactions with one another, a basin of attraction once again pulls some initial distributions away from equal division. Once we begin to think of correlation as endogenous, we will also want to explore the results of anticorrelation, for 2/3ers. Rational 2/3ers realize that their prospects are better when they avoid one another. If 1/2ers recognize each other with some measure of reliability in order to correlate, 2/3ers might deploy similar self-recognition ability to anticorrelate. When they can, the prospects get still worse for demand 1/2.

In Figure 4, 1/2ers and 2/3ers are pursuing opposing, symmetric strategies. Each has the same ability to recognize their own kind, and they have opposite preferences about whether to play with their own kind. Thus, both strategies are equally well positioned to pursue the strategy that benefits them. The correlation factor is still relatively small, to reflect difficulties in finding and identifying the individuals one prefers to play. The result of these changes is an increase in the size of the basin of attraction toward the polymorphism. Rough estimates show that the basin of attraction of the polymorphism is about 67% greater in Figure 4 than in Figure 3. (Increasing the degree of anticorrelation for 2/3ers only increases the size of the polymorphic

ability that S_i will play S_i is zero. This is analogous to the case of perfect positive correlation. In both cases if there are no players of the appropriate strategies left in the round, the first player in the pair dies off without playing.

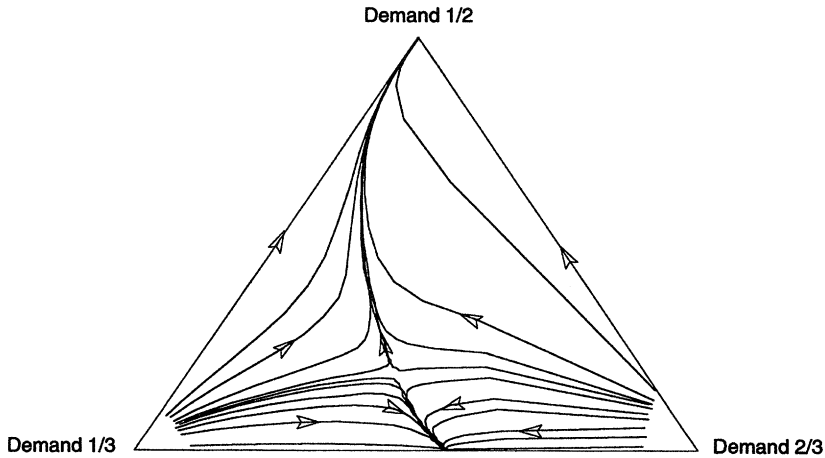

Figure 4. Correlations: 0, .2, −.2

basin of attraction, even when the degree of positive correlation for 1/2ers is commensurately increased.)

Because it treats demand 1/2 and demand 2/3 symmetrically, we think figure four is a move in the direction of realism. But what of the modest 1/3ers? Of course, as self-interested players, 1/3ers have no reason to correlate, nor to anticorrelate. 1/3ers always get their cake, no matter whom they play. They have no reason to expend time or energy targeting specific partner types. But surely that fact is itself an advantage which compensates them to some degree for their lower payoffs, and this should be reflected in the model. In order fully to reflect the advantages and disadvantages of correlated strategies, we should recognize the costs of being choosy.

In Figure 5 we have introduced a cost factor, c, to assign a cost to correlation and anticorrelation, as follows. For positive correlation, cost $= ce_i[1 - \mathrm{p}(\mathrm{S}_i)]$; for negative correlation, cost $= c|e_i|\mathrm{p}(\mathrm{S}_i)$; where c is the cost factor, e_i is the (positive or negative) correlation factor for S_i, and $\mathrm{p}(\mathrm{S}_i)$ is the relative frequency of the strategy in question within the population at that point in the round. These costs are then subtracted from each strategy's success at the end of each round, before the population size is renormalized. This is intended to account for the possibility that searching out specific strategies to play against, or declining to play the first individual one meets, could leave an individual without a partner in a given round. On our formulas, the costs for a given strategy are a function of how high a correlation (or anticorre-

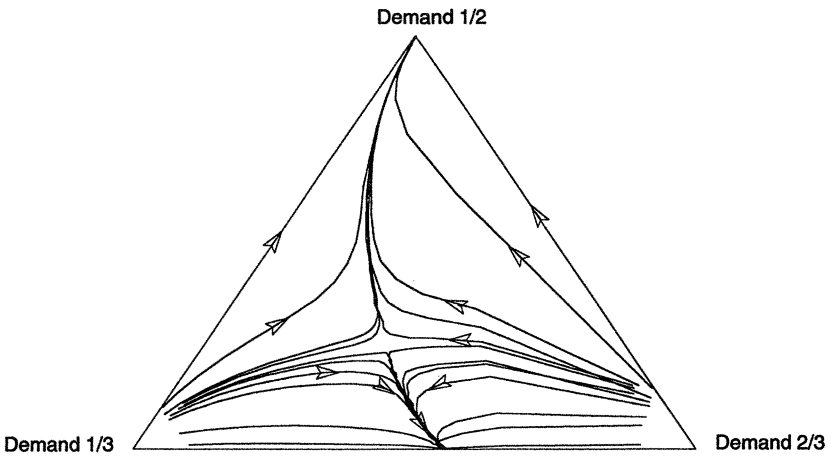

Figure 5. Correlations: 0, .2, −.2. Cost .3

lation) is attempted, and of how difficult it is to find partners of the desired sort. So, for instance, as the proportion of individuals playing your strategy drops, it becomes more difficult to find them—and this is reflected in a higher cost for those attempting to do so. The more determined you are to find them, the higher your correlation coefficient, and the more resources you can be expected to expend searching, on average.

Once again, the increase in realism increases the basin of attraction toward the greedy-modest polymorphism. Rough estimates again show that there is an increase in the size of the polymorphic basin of attraction (this time about 13%) in Figure 5 compared with Figure 4. Here c has been arbitrarily set at 0.3. As the value of c rises, the basin of attraction toward the polymorphism gets larger.

5. Conclusion. Typical applications of non-evolutionary game theory to rational decision theory seek to find the rational strategies for individuals in situations in which their decisions are mutually influencing. That endeavor is normative, rather than descriptive, in that it purports to tell us how it is *rational* to choose, given that others act rationally. Sometimes, however, such "optimizing models" are offered as evolutionary *explanations* of actual behavior.[26] In the study of non human

26. The relation between game theory's normative uses in decision theory and its explanatory uses in evolutionary theory is a complicated topic which we cannot treat in detail here. Sometimes, the idea is simply that a demonstration of the optimality of some behavior can be invoked as an explanation of the behavior, because natural selection finds (at least local) optima. But game theorists such as Skyrms model the evo-

animals, such models have been widely applied in behavioral ecology (Kamil, Krebs, and Pulliam 1987; Stephens and Krebs 1986), and less widely to social behavior (Emlen and Wrege 1994). And there is no principled reason why such models could not be usefully applied to human behavior.

We have suggested some broad guidelines for assessing explanations of human behavior which employ game theoretic models. We urged that such explanations should be representative, robust, and flexible. Brian Skyrms claims both robustness and flexibility for his account of the evolution of a propensity to demand 1/2 in divide-the-cake. We have argued, in effect, that his explanation does not display these virtues to the extent that he supposes.

Skyrms explicitly claims that the stability of the attracting equilibrium at demand 1/2 makes the details of the dynamics unimportant. But, in fact, the details of the dynamics determine the rationale for correlation, and hence the way in which it must be implemented in the model. Under the interpretations of the model which strike us as most realistic, there is a very considerable basin of attraction pulling the population toward a greedy-modest polymorphism in which demand 1/2 is wiped out. Thus, his explanation is neither as flexible nor as robust as it first appears.

Our point here is not to deny the possibility of good generalist evolutionary explanations of human behavior. Instead, we have sought to display the ambitions of such explanations, and the conditions they must meet to satisfy these ambitions. Generalist explanations are in place wherever, and because, circumstances are such that the specific causal details which produced the explanandum are not important for understanding it. Thus, they are best suited to the explanation of generalities which can be produced by a variety of distinct causal routes. If there are such generalities in human behavior, an explanation displaying how formal features of the behavior make it successful against the relevant range of alternatives in a variety of contexts will be important and insightful. But not every mathematical demonstration of some variety of robustness offers such a prospect.

REFERENCES

Alexander, Richard (1979), *Darwinism and Human Affairs.* Seattle: University of Washington Press.

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

---

lutionary process itself, where what evolves is not always rational. Thus, for example, Elliott Sober has pointed out to us that cooperation in the one-shot prisoner's dilemma can evolve when there is correlation among interactors, even though it is still rational to defect.

Batterman, Robert (1992), "Explanatory Instability", *Noûs* 26: 325–348.
Boyd, Robert and Peter J. Richerson (1985), *Culture and the Evolutionary Process.* Chicago: University of Chicago Press.
Buss, David (1994), *The Evolution of Desire.* New York: Basic Books.
Campbell, Donald T. (1975), "On the Conflicts Between Biological and Social Evolution and Between Psychology and Moral Tradition", *American Psychologist* 30: 1103–1126.
Cosmides, Leda and John Tooby (1992), "Cognitive Adaptations for Social Exchange", in Jerome Barkow, Leda Cosmides, and John Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* New York: Oxford University Press, pp. 163–228.
Daly, Martin and Margo Wilson (1988), *Homicide.* New York: Aldine de Gruyter.
D'Arms, Justin (1996), "Sex, Fairness, and the Theory of Games", *Journal Of Philosophy* 93: 615–627.
Dawkins, Richard (1976), *The Selfish Gene.* Oxford: Oxford University Press.
———. (1982), *The Extended Phenotype.* New York: Oxford University Press.
Emlen, Stephen T. and Peter H. Wrege (1994), "Gender, status and family fortunes in the white-fronted bee-eater", *Nature* 367, 13: 129–132.
Fodor, Jerry (1983), *The Modularity of Mind.* Cambridge, MA: MIT Press.
Gibbard, Allan (1982), "Human Evolution and the Sense of Justice", *Midwest Studies in Philosophy* VII: 31–46.
Gould, Stephen J. and Richard W. Lewontin (1979), "The Spandrels of San Marco and the Panglossian Paradigm: a critique of the adaptationist programme", *Proceedings of the Royal Society of London*, B 205: 581–598.
Kamil, Allan C., John R. Krebs, and H. Ronald Pulliam (eds.) (1987), *Foraging Behavior.* New York: Plenum.
Kitcher, Philip (1985), *Vaulting Ambition.* Cambridge, MA: MIT Press.
Maynard Smith, John (1982), *Evolution and the Theory of Games.* New York: Cambridge University Press.
Maynard Smith, John and G.R. Price (1973), "The Logic of Animal Conflict", *Nature* 146: 15–18.
Nash, John (1950), "The Bargaining Problem", *Econometrica* 18: 155–162.
Pinker, Stephen (1994), *The Language Instinct.* New York: HarperCollins.
Pinker, Stephen and Paul Bloom (1992), "Natural Language and Natural Selection", in Jerome Barkow, Leda Cosmides, and John Tooby (eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture.* New York: Oxford University Press, pp. 451–494.
Railton, Peter (1978), "A Deductive-Nomological Model of Probabilistic Explanation", *Philosophy of Science* 45: 206–226.
———. (1981), "Probability, Explanation, and Information", *Synthese* 48: 233–256.
Skyrms, Brian (1994), "Sex and Justice", *The Journal of Philosophy* 91: 305–320.
———. (1996a), *Evolution of the Social Contract.* New York: Cambridge University Press.
———. (1996b), *The Social Contract Naturalized.* MBS 96–31, Institute for Mathematical Behavioral Sciences Technical Report Series. Irvine: University of California, Irvine.
Smale, S. (1980), "What is Global Analysis?", in *The Mathematics of Time: Essays on Dynamical Systems, Economic Processes, and Related Topics.* New York: Springer-Verlag, pp. 84–89.
Sober, Elliott (1993), *Philosophy of Biology.* Boulder, CO: Westview Press.
Stephens, D.W. and J.R. Krebs (1986), *Foraging Theory.* Princeton: Princeton University Press.
Sterelny, Kim (1992), "Evolutionary Explanations of Human Behavior", *Australian Journal of Philosophy* 70, 2 (June): 156–173.
———. (1995), "Review of 'The Adapted Mind'", *Biology and Philosophy.* 10, 3: 365–380.
Symons, Donald (1979), *The Evolution of Human Sexuality.* Oxford: Oxford University Press.
Tooby, John and Leda Cosmides (1992), "The Psychological Foundations of Culture", in Jerome Barkow, Leda Cosmides, and John Tooby (eds.), *The Adapted Mind:*

Evolutionary Psychology and the Generation of Culture. New York: Oxford University Press, 19–136.
Turke, Paul and Laura Betzig (1985), "Those who can, do: Wealth, Status, and Reproductive Success on Ifaluk", *Ethology and Sociobiology* 6: 79–87.

APPENDIX

As we note in the paper, our model differs from Skyrms's in a number of respects. First, our populations are finite. Second, we pair players individually in a given round until the population existing in that round is exhausted. This allows us to model a situation in which each strategy is either positively or negatively correlated and where each such correlation is unconstrained by the correlations being pursued by the other strategies. These differences, we believe, are justified because they allow for a more realistic account of the evolutionary problem. Furthermore, despite these differences, it should be emphasized that our model completely reproduces Skyrms's results in the situations he considers.

Let S_1, S_2, S_3 represent respectively 1/3ers, 1/2ers, 2/3ers; that is, they label the different strategies. Let N_1, N_2, N_3 represent the number of the various types in the population. These numbers are input at the beginning of the game. We have:

$$N_1 + N_2 + N_3 = N_{tot}.$$

Hence, the proportion (or relative frequency) of the strategies, $\Pr(S_i)$ ($i \in \{1,2,3\}$), in the beginning of our simulation is:

$$\Pr(S_i) = \frac{N_i}{N_{tot}}.$$

Likewise, before the start of the game we input the various correlation factors. These are numbers e_i. (For example, in Figure 4 we have $e_1 = 0$, $e_2 = .2$, $e_3 = -0.2$.) The round begins as follows: We choose a random number between 0 and 1. This interval is divided into three segments the *lengths* of which are the relative frequencies $\Pr(S_i)$. Suppose for this example that the population contains some 1/2ers ($N_2 \neq 0$) and the random number lies in the interval of length $\Pr(S_2)$. This means that a 1/2er has been chosen to play first. Next we need to calculate the various conditional probabilities that that player will play a 1/3er, a 1/2er, and a 2/3er.

Because our model samples without replacement we need first to recalculate the relative frequencies of the various players given that we have selected a 1/2er first. These are given as follows:

$$\Pr(S_1) = \frac{N_1}{N_{tot}-1}, \ \Pr(S_2) = \frac{N_2-1}{N_{tot}-1}, \ \Pr(S_3) = \frac{N_3}{N_{tot}-1}.$$

The probability that our 1/2er will meet another 1/2er is then determined as follows. (The correlation coefficient e_2 was input at the beginning of the game.)

If $e_2 \geq 0$, then

$$\Pr(S_2|S_2) = \Pr(S_2) + e_2 \Pr(not - S_2).$$

If $e_2 < 0$ then

$$\Pr(S_2|S_2) = \Pr(S_2) + e_2 \Pr(S_2).$$

The probability that she will meet a 1/3er is given by

$$\Pr(S_1|S_2) = (1 - \Pr(S_2|S_2))\frac{\Pr(S_1)}{\Pr(S_1)+\Pr(S_3)};$$

and the probability that she will meet a 2/3er by

$$\Pr(S_3|S_2) = (1 - \Pr(S_2|S_2))\frac{\Pr(S_3)}{\Pr(S_1)+\Pr(S_3)}.$$

In general, correlations are input to yield $\Pr(S_i|S_i)$; $i \in \{1,2,3\}$. For $j \neq k \neq i$ we have

$$\Pr(S_j|S_i) = (1 - \Pr(S_i|S_i))\frac{\Pr(S_j)}{\Pr(S_j)+\Pr(S_k)};$$

We now divide the unit interval into 3 segments of length $\Pr(S_2|S_2)$, $\Pr(S_1|S_2)$, and $\Pr(S_3|S_2)$ and choose another random number n in that interval. Which of these segments of [0,1] n finds itself in determines who our 1/2er plays with. Suppose that n is in the segment corresponding to the strategy S_1. Then our 1/2er plays a 1/3er and since together they demand less than 100% of the cake, they each get what they ask for. They survive to play again in the next round. On the other hand, if n is such that our 1/2er must play a 2/3er, then since together they demand more than 100% of the cake, neither player receives a share. In this case neither survives to play again in the next round.

The round continues in exactly the same way. (i) A player is randomly chosen from the remaining population. (ii) The relative proportions of the different strategies are recalculated. (iii) The various conditional probabilities for the chosen player to play a 1/3er, a 1/2er, and a 2/3er are then calculated. (iv) A second random number is chosen which then determines according to the probabilities in (iii) who the first player meets. (v) Individuals who have received a payoff survive and proceed to the next round; those who have not are killed off.

Finally, at the end of the round–when there are no more players to be chosen—we renormalize the populations so that the total number is once again N_{tot} and the different strategies S_i are represented in new relative proportions

determined by the payoffs they received in the last round: Let N_{fi} be the number of players of strategy S_i remaining at the end of the round. Let $N_{final} = 1/3N_{f1} + 1/2N_{f2} + 2/3\ N_{f3}$. Finally, define N_{newi} to be the number of players of strategy S_i starting the next round. We have

$$N_{new1} = 1/3N_{f1}\frac{N_{tot}}{N_{final}}$$

$$N_{new2} = 1/2N_{f2}\frac{N_{tot}}{N_{final}}$$

$$N_{new3} = 2/3N_{f3}\frac{N_{tot}}{N_{final}};$$

where, of course,

$$N_{tot} = \sum_i N_{newi}.$$