Philosophy and Phenomenological Research
Vol. LIX, No. 1, March 1999

Models and Reality—A Review of Brian Skyrms's *Evolution of the Social Contract*

MARTIN BARRETT, ELLERY EELLS, BRANDEN FITELSON, AND ELLIOTT SOBER*
University of Wisconsin, Madison

Human beings are peculiar. In laboratory experiments, they often cooperate in one-shot prisoners' dilemmas, they frequently offer 1/2 and reject low offers in the ultimatum game, and they often bid 1/2 in the game of divide-the-cake. All these behaviors are puzzling from the point of view of game theory. The first two are irrational, if utility is measured in a certain way.[1] The last isn't positively irrational, but it is no more rational than other possible actions, since there are infinitely many other Nash equilibria besides the one in which both players bid 1/2. At the same time, these behaviors seem to indicate that people are sometimes inclined to be cooperative, fair, and just. In his stimulating new book, Brian Skyrms sets himself the task of showing why these inclinations evolved, or how they might have evolved, under the pressure of natural selection. The goal is not to justify our ethical intuitions, but to explain why we have them.[2]

Skyrms's strategy is to equate "our sense of justice" with a set of behavioral dispositions. He assumes that these dispositions evolved because they led to behaviors that were on average fitter than the ones triggered by alterna-

* We thank Steven Orzack and Larry Samuelson for useful comments.

[1] Skyrms explicitly notes the need for the "if" in this statement; there is nothing in game theory that requires that utility be measured in terms of material resources. For example, if people care enough about others, then it isn't irrational for them to cooperate in situations that may seem to be one-shot prisoners' dilemmas. However, if games are defined in part by the utilities involved, then a game in which cooperation is the rational action is not a one-shot prisoners' dilemma. Perhaps, then, we should interpret Skyrms' models, not as explaining why people sometimes behave irrationally, but as explaining why people sometimes perceive utility in ways that are orthogonal to material self-interest.

[2] Although Skyrms does a good job describing mathematical ideas informally, those not already familiar with game theory may find it useful to consider simple algebraic representations of some of the arguments that Skyrms presents. See, for example, Sober (1994, p. 17) for a presentation of R.A. Fisher's analysis of sex ratio evolution and Sober (1994, pp. 137–39) for the calculation of the equilibrium in the Hawk-Dove game.

tive dispositions. In the three games mentioned above, Skyrms invokes the concept of positive correlation between individuals.[3] For example, although defectors do better than cooperators in one-shot prisoners' dilemmas when players pair at random, the reverse is true when there is a sufficiently strong tendency for like to pair with like.[4]

Skyrms's framework has a surprising consequence—it entails that modern human beings are inclined to behave justly. This doesn't mean that we all act justly all the time, anymore than a bunch of soluble sugar cubes must all dissolve. But the disposition is in us all. Is this implication something we should accept? The idea that some people not only fail to act justly, but also lack the inclination to do so, cannot be dismissed out of hand. And even if we assume that all modern humans have a "sense of justice," there is the issue of how well this sense is captured by the disposition to bid 1/2 in divide-the-cake. The problem begins in our own society. We suspect that many people in the here and now would disagree with the idea that equal division of a resource is always what justice requires—especially if one party would benefit from the resource more, or if one had invested more in producing it. These common intuitions are used by utilitarians, Rawlsians, and libertarians to erect their separate theories, none of which entails that resources should always be split 50/50. An additional dimension of this problem comes to light when we recognize that justice has meant different things in different places and times. How universal is the sense of justice that Skyrms takes as his *explanandum*? Perhaps there are core ideas about justice that are cultural universals. Skyrms offers no evidence on this point. In fact, the empirical information mustered in Skyrms's book is extremely slender. He cites the experiments we mentioned. After that, his efforts go into formulating and exploring mathematical models, whose intended results are that they predict the behaviors observed in the laboratory.

There is an alternative explanation that is worth exploring for some of the behavioral tendencies that Skyrms considers. The Newcomb problem and the problem posed by Fisher's smoking hypothesis structurally resemble the one-shot prisoners' dilemma (Lewis 1979). A standard diagnosis of why people reach the wrong decision in Newcomb/Fisher settings is that they confuse cause and correlation. Perhaps people do the same thing in the one-shot prisoners' dilemma and in divide-the-cake as well; they say to themselves "the other player is just like me, so whatever I do, he'll do too." If so, the tendency to cooperate in the prisoners' dilemma and to bid 1/2 in divide-the-cake has nothing specifically to do with ethics.

[3] This idea also is used in the section of the chapter on meaning called "Signals for Altruists."

[4] Here Skyrms employs an idea that has figured prominently in discussion of the evolution of altruism; in addition to the references that Skyrms supplies, see Sober and Wilson (1998) for a review of this work.

Skyrms begins his book by drawing an analogy between sex ratio evolution and the game of divide-the-cake. In fact, there are important structural dissimilarities here. Skyrms considers only pure strategies in his treatment of divide-the-cake, and to explain why Fair evolves rather than a Modest-Greedy polymorphism, he invokes an assumption of positive correlation. His models involve deterministic dynamics, reflecting the assumption of infinite population size. In the case of sex ratio evolution, however, mixed strategies must be considered from the start; a parental pair's mix of sons and daughters is the result of its producing a son with probability p and a daughter with probability (1-p). In Fisher's model, a 1:1 population sex ratio evolves because parental pairs form at random. W.D. Hamilton showed how patterns of inbreeding and dispersal can make uneven sex ratios evolve. To explain, in the case of random mating, why an even sex ratio is achieved in a population by having all parents produce sons and daughters with equal probability, rather than by having different parents pursue different mixed strategies, finite population size must be invoked (on which more below); however, when an uneven sex ratio evolves in a structured population, finite population size isn't needed to explain why all individuals follow the same mixed strategy (Orzack *et al.*, 1991).

As mentioned earlier, the key idea that Skyrms invokes is that of positive correlation. Skyrms defends the realism of this assumption; if interactions tend to be among relatives, then interactors will tend to resemble each other, if the phenotype in question is influenced by genes. How far does this observation take us in establishing the plausibility of Skyrms's proposed explanations? An additional question that needs to be addressed is whether the different strategies that Skyrms describes are heritable. Do offspring tend to resemble their parents? This might be due to shared genes or to learning and imitation. Since subjects did not all act the same in the experiments that Skyrms cites, the question of heritability (at least in the present, if not in the unobservable past) should be tractable. But more importantly, we need to ask whether our ancestors (recent or more ancient) really played divide-the-cake, the ultimatum game, and the one-shot prisoners' dilemma. A selective explanation of the sex ratio strategy *now* found in a species depends on the fact that sex ratio *was* an adaptive problem when the species was evolving. The symmetrical point is that Skyrms is committed to the idea that human beings now bid 1/2 when they play divide-the-cake because earlier human populations faced adaptive problems that had the structure of divide-the-cake. Some doubts arise here (D'Arms, 1996; D'Arms, Batterman and Gorny, 1998). Although it makes sense to suppose that our ancestors had to divide resources, how plausible is it to think that they had to bid simultaneously and that they would have lost everything if their bids had totaled more than 100%? As for one-shot prisoners' dilemmas, the fact that ancestral humans lived in small bands suggests that interactions were almost never one-shot.[5] An analogous limitation attaches to the ultimatum game (sometimes called "take it or leave it"), which requires that the proposer has no reputation to maintain. The ecological validity of all three games is highly questionable.

When Skyrms shows that positive correlation helps Fair evolve in divide-the-cake, his point is that the more positive correlation there is among interactors, the larger the basin of attraction is of the Fair monomorphism, compared with the Modest-Greedy polymorphism. Skyrms views the robustness of 100% Fair as evidence that it is more "likely" to have evolved; it is the end state of a "larger" set of initial conditions. However, the question of robustness can be raised in other ways. For example, Skyrms considers positive correlation but not negative correlation among interactors; D'Arms, Batterman, and Gorny (1998) have found that certain types of anti-correlation lead to some fairly robust polymorphisms. They consider patterns of association in which individuals tend to pair with those unlike themselves, irrespective of whether this does them any good. The realism of this arrangement may be doubted. However, suppose that individuals seek out partners with an eye to maximizing their own advantage.[6] Then Greedy will want to pair with Modest, but not with itself or with Fair. Fair will want to pair either with itself or with Modest, but not with Greedy. And Modest will be indifferent. We suspect that this pattern of association qualitatively resembles the one that D'Arms and Batterman considered—it should provide a larger basin of attraction for the polymorphic equilibrium.

The question of robustness also should be posed by adding mixed strategies to the three pure strategies that Skyrms considers. For example, consider the strategy Mix, which has the agent bid 1/3 with probability 0.5 and bid 2/3 with probability 0.5. Here are the payoffs to row:

	Modest	Fair	Greedy	Mix
Modest	1/3	1/3	1/3	1/3
Fair	1/2	1/2	0	1/4
Greedy	2/3	0	0	1/3
Mix	1/2	1/6	1/6	1/3

In this game, Mix and Fair are both evolutionarily stable strategies, and the Modest-Greedy polymorphism is an evolutionarily stable state. The main diagonal of this table indicates that Fair will evolve if there is a sufficiently

[5] It is worth noting that cooperative strategies (such as tit-for-tat) can evolve in iterated prisoners' dilemmas when individuals pair at random. Positive correlation isn't essential.

[6] This is a pervasive fact about human (and some nonhuman) interactions—we often *choose* the individuals with whom we associate; the pattern of association is not exogenously fixed. The consequences this has for the evolution of altruism are discussed in Sober and Wilson (1998).

strong positive correlation among interactors. In this respect, the competition among these four strategies resembles the competition among the three pure strategies that Skyrms considers. The situation becomes more complicated if we allow finite population size to introduce an element of drift into the dynamics; in this circumstance, a monomorphic mixed strategy is more robust than a polymorphism of pure strategies (Hines and Anfossi 1990, Young 1993)—the idea of positive correlation isn't needed to explain why Mix is more likely to evolve than a Modest-Greedy polymorphism. Unfortunately, this fact does not settle whether Fair is more robust than Mix; we can offer no solution to this problem at present. Mix is just an example of the mixed strategies that need to be considered before Fair can be regarded as a robust solution to divide-the-cake.

Our technical comments in this review should not obscure what we take to be the main point. Just as in the case of sex ratio evolution, mathematical models are one thing, empirical reality another. We are not complaining about idealization. Rather, our claim is that a lot more empirical work is needed to show how model and reality are connected. Until then, it is an open question how much correlated evolutionary game theory in divide-the-cake, the one-shot prisoners' dilemma, and the ultimatum game has to tell us about the evolution of the social contract.

References

D'Arms, J. (1996), "Sex, Fairness, and the Theory of Games." *Journal of Philosophy* 96: 615–27.

D'Arms, J., Batterman, R., and Gorny, K. (1998), "Game Theoretic Explanations and the Evolution of Justice," *Philosophy of Science* 65: 76–102.

Hines, W. and Anfossi. D. (1990), "A Discussion of Evolutionarily Stable Strategies", in S. Lessard (ed.), *Mathematical and Statistical Developments of Evolutionary Theory*, Dordrecht: Kluwer, pages 229–67.

Lewis, D. (1979), "Prisoners' Dilemma is a Newcomb Problem." *Philosophy and Public Affairs* 8: 235–40.

Orzack, S., Parker, E., and Gladstone, J. (1991), "The Comparative Biology of Genetic Variation for Conditional Sex Ratio Adjustment in a Parasitic Wasp, *Nasonia vitripennis*." *Genetics* 127: 583–99.

Sober, E. (1993), *Philosophy of Biology*. Boulder, Colorado: Westview Press.

Sober, E. and Wilson, D. (1998), *Unto Others—the Evolution and Psychology of Unselfish Behavior*. Cambridge: Harvard University Press.

Young, P. (1993), "An Evolutionary Model of Bargaining". *Journal of Economic Theory* 59: 145–68.