Artificial Justice

J. McKenzie Alexander
Logic & Philosophy of Science
University of California, Irvine
Irvine, CA 96297

Abstract

Recently, there has been an attempt among some philosophers to provide cultural evolutionary grounds for certain norms of distributive justice. The most noteworthy attempt (Skyrms, 1996) uses a simple evolutionary model based upon the replicator dynamics. I argue that the replicator dynamics is not the most appropriate model of interactive human behavior in societies, and present an alternative agent-based model. I demonstrate how the conditions under which norms of distributive justice arise depend on how we construe the underlying evolutionary dynamics.

0. Introduction

The presence of a concept of justice is a fundamental property of human societies. Intuitions over what exactly this concept includes vary widely, but, generally speaking, a concept of justice can be thought of as a set of principles identifying the appropriate relationship between individuals and the society to which they belong. Two topics of natural interest to philosophers concern *descriptive* and *normative* questions of the concept of justice: What concept of justice do people actually have? Why should people hold one concept of justice instead of another?

Constructing a theory of justice which is both descriptively and normatively adequate is a daunting task, given that it must accurately describe the concept of justice individuals have, it explains why individuals should have *that* particular concept of justice instead of any other, and it explains what we mean when we say we *ought* to follow certain principles of justice. In this paper I shall concentrate on a very minor question about distributive justice which properly belongs to a general theory of justice. This concentration will not prove problematic for my purposes, though, as presently I wish to argue that philosophers interested in the sorts of questions mentioned above can benefit by approaching these questions employing techniques used within the artificial life community. To argue this, it suffices to show how certain fundamental questions of distributive justice can be profitably studied within the artificial life framework. It must be remembered that the models discussed here address only a very small area of the full concept of justice most people possess; a complete account of our concept of justice would need to expand upon the treatment given here to arrive at a more empirically adequate account.

1. The distribution problem

In its most general form, the distribution problem consists of a set of goods G to be distributed among the members of a population P, with situational considerations S, subject to two constraints:

1. No good is assigned to two members of the population.[1]

2. Each good is assigned to some member of the population.[2]

[1]It may seem that this restriction limits the generality of distribution problems considered, but it really does not. To handle problems involving public goods which may be shared among people, such as highways, parks, and baseball fields, we simply need to adjust our conception of the good to be distributed. Instead of conceiving of the good (say, the highway) as a single item which must be assigned to one and only one agent, we do not assign the highway itself but rather *time-shares* in sections of the highway, one time-share to each person who desires a part of it. After all, although the highway, as a whole, may be used by more than one person at the same time, no single part of the highway (one hopes) will be used by more than one person at the same time. This approach allows us to make the simplifying assumption that no good can be assigned to two members of the population without a loss of generality.

[2]This restriction simply requires any solution to be Pareto-optimal. This captures the commonly held belief that, if the position of any person may be improved without negatively affecting the position of anyone else, then that person's position *should* be so improved. Of course, if the set G we are distributing over the population contains undesirable items (say, diseases, debts, or unpleasant duties), it might be appropriate to drop this requirement.

The situational considerations S allow us to incorporate various relevant facts concerning needs, rights, prior claims to some of the goods in G, and so on, into the distribution problem.[3]

A *solution* to the distribution problem is an assignment of sets of goods to each member of the population subject to the above constraints. (By a *good* I mean any object of value which is capable of being assigned to any arbitrary member of the population. The point of this restriction is to eliminate from consideration those objects judged to be of value which cannot be, for whatever reason, objects of exchange.) Clearly the "problem" here is not with finding a solution, for many solutions exist: any function $f_S : G \to P$ gives a solution satisfying our two constraints. The problem concerns *selecting* a particular solution (or solutions) out of the many possible ones available to receive the label "fair" or "just."

Two particular distribution problems are of particular philosophical interest because their simple structures seem to correspond to primitive principles, or norms, of distributive justice. These distribution problems have the additional virtue of being extensively studied by economists, resulting in a large body of theoretical and empirical results which may be brought to bear on the problem. In the rest of this section, I shall describe the two problems which shall receive our attention for the rest of this paper, with brief summaries of some of the relevant experimental work on the subject. Throughout I endeavor to explain why these particular problems deserve our attention, their simple forms notwithstanding.

The Nash bargaining game

The simplest distribution problem of interest involves allocating a divisible good G, which I shall refer to as "cake," between two people.[4] For simplicity, we assume the measure of the amount of cake is such that the total amount of cake available for distribution is 10, where the units of measurement correspond to a natural quantity, e.g., slices of cake. Furthermore, let us assume that the two individuals are perfectly symmetric in all relevant respects. This assumption insures that the good is equally useful to each person, each person has the same need, and that no person has a prior claim on the good that would trump the other person's claim, among other things.

In this symmetric case, we have strong intuitions urging that the "just" or "fair" distribution allocates exactly half of the cake to each person. Our common intuitions suggest that the relevant principle of distributive justice almost everyone holds is the one singling out the equal split as the (uniquely) just solution. Although few would object to saying that the equal split is the correct principle of distributive justice to hold, at least for the perfectly symmetric case of divide-the-cake, studies indicate that this intuition is, in fact, widely shared.

In 1974, Nydegger and Owen conducted an experimental test of people's behavior for the game of divide-the-cake (although they had subjects divide a dollar instead of a cake). Not surprisingly, they found that *all* pairs of subjects agreed on the 50–50 split. While some doubt over the generality of the results may be warranted given the small, biased sample size, their claim seems correct that, "The outcome of this study is quite impressive if for no other reason than the consistency of its results." (Nydegger and Owen, 1974, page 244) seems correct.

That people *do* ask for half of the cake in a perfectly symmetric situation is indisputable. Explaining *why* people always ask for half of the cake is more difficult. The equal split in the game of divide-the-cake is an equilibrium in informed, rational self-interest (also known as a Nash equilibrium) in that each player's request is optimal given the other player's request.[5] However, given the particular structure of divide-the-cake, *every* solution which does not give all of the cake to one player is a Nash equilibrium. Yet the common game-theoretic solution concept of a Nash equilibrium does not help us identify *why* the equal-split ought to be favored over any other alternative.

It should be noted that Nash (1950) presents an argument which singles out, in certain cases, the equal-split outcome from the many other Nash equilibria possible. Chronicling the objections made against, and virtues of, Nash's approach would take me too far afield; however, the interested reader may consult Luce and Raiffa (1957, pp. 128–134) for an excellent criti-

[3] For a discussion of the effects situational considerations have on the favored solution to particular distribution problems, see Yaari and Bar-Hillel (1984).

[4] The one-person distribution problem has an obvious solution—the person receives everything—that I take to be unproblematic. Conceivably, though, one could argue that if the person were incapable of using all of the good, giving her more than she could use would be unjust since it eliminates the possibility of another individual using the remainder of the good. In all of the cases of the distribution problem I consider, these sorts of complicating contextual factors are assumed to be impossible, simply because the additional complexity introduced obscures the basic problem.

[5] If both A and B ask for half of the cake, neither player can improve her situation by changing her request, provided the other player's request remains fixed. For example, if A decides to ask for 60% of the cake when B still requests 50%, the total amount they ask for overshoots the amount of cake available, and each player receives nothing.

cism of Nash's approach. Skyrms (1996) contains a good discussion on why explaining the equal-split outcome of divide-the-cake should be so difficult.

One might suspect that the difficulty in explaining the equal-split outcome is primarily due to certain artificialities in our framing of the problem: since agents only play the game once, there is no opportunity for them to communicate their preferences to the other player. In an *iterated* distribution problem, the iteration creates a kind of communication between the agents, allowing one agent to, in effect, tell the other that she will not accept an offer below a certain threshold, possibly allowing the other to coordinate on the "fair" solution. Consider the following variation of the Nash game: as before, the two players must agree on how to divide a cake but, unlike before, we do not require them to arrive at an agreement by the end of the first round of play. At any particular time it is one player's turn to suggest a possible division of the cake. If the other player accepts the proposed division, the game ends; if the other player does not accept the proposal, then the two swap roles and the other player may suggest a division. This alternation continues until a decision is reached. The catch is that the total amount of cake available for dividing decreases as time passes—after t minutes, only $100\delta^t\%$ remains, for some $\delta \in [0, 1]$.

In this new setting, there are still an infinite number of Nash equilibria, but the majority of them fail to be subgame perfect. Rubenstein (1982) showed that if we require the agents to follow subgame perfect strategies, there is a single equilibrium in which the first player proposes slightly more than half of the cake and the second player agrees immediately to this distribution. Initially, this result may seem to settle the question as to why people tend to ask for half in divide-the-cake type games, but closer examination reveals that the matter is not so simple. Kreps (1990) notes that if the cake is not infinitely divisible, the problem of multiple equilibria reappears. Moreover, if the two players differ in their response time, the subgame-perfect equilibrium selected distributes the good proportionally to the response rate of individual agents; and if each agent incurs a cost when making a proposal, then the agent incurring the smallest cost receives the majority of the cake. Since all three phenomena (discrete granularity of the good, variable response time, and variable cost of interaction) are present in real bargaining situations, we see that considering the iterated game as a way to settle the question of why people ask for half of the cake does not suffice.

The ultimatum game

A slightly more complicated distribution problem arises when we lift the requirement (present in divide-the-cake) that the two agents are perfectly symmetric. The *ultimatum* game provides a useful example of such a game. Here we again have two agents A and B who need to determine how best to share some good (say a cake) between them. In the ultimatum game we assume that one party, say A, has initial possession of the cake, and presents B with an offer of how much of the cake A is willing to give B (this is the ultimatum). B may either accept or reject the offer. If B rejects the offer, A and B each receive nothing (if you want a reason for this, suppose they begin arguing and the cake spoils). If B accepts the offer, each person receives the appropriate amount of cake. The fact that each player in this game has distinct (and different) roles makes the extended-form representation shown in figure 1 the most natural one.

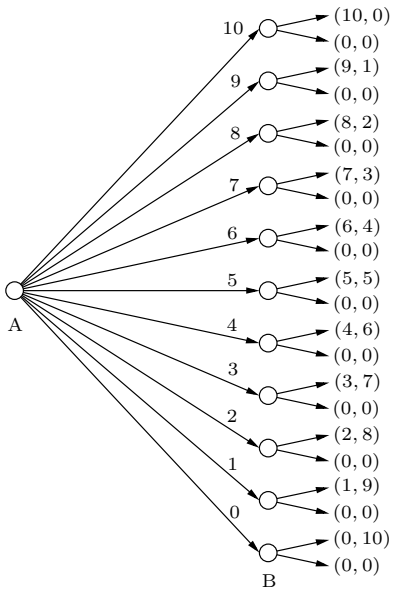

Figure 1: Extended form representation of the ultimatum game

Intuitions as to the "just" resolution to this particular distribution problem seem to vary more than for the Nash bargaining game. A fairly strong intuition exists which says that, since neither player has any special need or entitlement to the cake (the fact that A has possession of the cake is seen as a historical accident), the equal split, or at least something relatively close to the equal split, is again perceived as the only just outcome.

Between 1982 and 1990, many experiments were con-

ducted to determine people's actual behavior in the ultimatum game, as the original results reported by Güth (1982) were quite surprising, reporting that people's actual behavior did not conform to the game-theoretic predictions. (Traditional game theory predicts player B will accept any nonzero offer; after all, some of the cake is better than none of the cake, and so as long as player A leaves player B with some cake, B should take it.) Although some people behaved in accordance with the "unjust" solution predicted by traditional game theory, the modal offer was the equal-split. Numerous follow-up experiments attempted to isolate the factors eliciting this response.[6] Ultimately, though, the experimental evidence suggests that there exists a principle of distributive justice, guiding people's behavior in the ultimatum game, which favors the equal split (or something close to it).

3. Evolutionary Explanations of the Equal Split

The Nash bargaining game

Skyrms (1996) offers an evolutionary explanation of why the equal split may be so widely followed, using the replicator dynamics of Taylor and Jonker (1978). As detailed discussions of his model may be found elsewhere (Skyrms, 1996), I shall only sketch the details here. For simplicity, assume that the cake is sliced into 10 pieces, and that individual requests are restricted to an integer number of slices. We represent the total state of the population by a vector $\vec{s} = (s_0, s_1, \ldots, s_{10})$ where s_i denotes the proportion of the population asking for i slices. Assume that the current "growth" rate of the types of individuals asking for i slices of the cake is approximately equal to the expected fitness of the demand i request in the population $\vec{s}$ [denote this by $F(i|\vec{s})$]. According to the replicator dynamics, the rate of change of the proportion of individuals requesting i slices of the cake is given by $\frac{ds_i}{dt} = s_i\big(F(i|\vec{s}) - F(\vec{s}|\vec{s})\big)$, where $F(\vec{s}|\vec{s})$ denotes the average fitness of the population. One should note that we are describing a cultural evolutionary process here, where the evolutionary dynamics describe changes in *beliefs*, rather than an evolutionary process operating on the biological level.

In a series of 100,000 trials, each trial beginning from a randomly selected point in the state space of the population, the final (converged) state of the population is distributed as follows:

Polymorphism	Count
Fair division	60822
4–6	28131
3–7	9324
2–8	1663
1–9	60
0–10	0

Notice that the population evolves to a state where fair division is the predominant norm approximately 60% of the time. The other rows in the table correspond to polymorphic states of the population where some fraction of the population requests i slices of the cake and the remainder requests $10 - i$ slices.

This provides the start of an evolutionary explanation of the norm of fair division in the Nash bargaining game, but it has the unfortunate consequence that it depends, rather heavily, on the initial conditions of the population. In terms of an explanation of why we think the equal split is *just*, this account appears a bit wanting. Although it offers some explanation of why the norm of fair division proves so widespread (namely, the initial conditions of our population lay within the basin of attraction for fair division), it does not explain why we think we *ought* to ask for half of the cake: had the initial conditions been otherwise, the evolutionary dynamics would have carried the population to (say) the 4–6 polymorphism, and people's beliefs as to what the appropriate sort of division was would be very different.

Skyrms does show that if interaction in the model is correlated (that is, people are more likely to interact with people following the same strategy than another), then once the degree of correlation exceeds a certain value, the basin of attraction for fair division expands to the interior of the state space. This seems to provide a plausible beginning to an evolutionary explanation of (certain) norms of distributive justice.

The ultimatum game

Skyrms also constructs a replicator dynamic model of the ultimatum game in order to see whether a similar evolutionary explanation of the norm of fair division existing in that game can be provided as well. In order to make such a model tractable, one needs to restrict the set of possible strategies, for even if we assume that the cake divides into only ten slices, there are $11 \cdot 2^{11}$ possible strategies an individual may follow.[7] One particularly interesting group of strategies to study is as follows:

[6] See, for example, Binmore *et al.* (1985), Güth and Tietz (1985), Neelin *et al.* (1988), Ochs and Roth (1989), Roth *et al.* (1991). For comprehensive surveys of the relevant experimental results, see Thaler (1988), Roth (1995), and Güth and Tietz (1990).

[7] Each strategy consists of two parts: the amount of cake one offers when one has possession of the cake, and the offers one is willing to accept. There are 11 possible offers one may make (offer 0 slices,..., offer 10 slices) and 2^{11} possible acceptance strategies.

Strategy	Offer	Accept
Gamesman	1	anything
S2	1	nothing
S3	1	accept 5, reject 1
Mad Dog	1	accept 1, reject 5
Easy Rider	5	anything
S6	5	nothing
Fairman	5	accept 5, reject 1
S8	5	reject 5, accept 1

Not every initial state of the population converges to a state where a "fair" or "just" strategy (Fairman or Easy Rider) dominates. A population in which all strategies appear equally likely converges to a state containing (roughly) 87% Gamesman and 13% Mad Dogs (Skyrms, 1996, pg. 31). However, certain initial population proportions *do* lead to states where only the strategies of Fairman and Easy Rider are present. This account still falls prey to the previous criticism of unacceptable dependence on the initial conditions.

Criticisms of Skyrms's model

Several criticisms of Skyrms's project of providing an evolutionary account of distributive justice have appeared in the philosophical literature. Since covering these criticisms in detail would take me too far from my present purpose, the interested reader may wish to consult Barrett (1999), Kitcher (1999), and D'Arms *et al.* (1998) for further discussion. I shall concentrate here on two criticisms:

1. The appropriateness of using the replicator dynamics to model human populations.

2. Concerns regarding Skyrms's introduction of correlation in his model of the Nash bargaining game.

First, using the replicator dynamics to model human populations requires that one make two assumptions which, taken together, seem highly implausible. When deriving the replicator dynamics, one needs to assume that the size of the population is sufficiently large to warrant identifying individual fitness with expected fitness. (This allows one to keep track of the evolution of the proportions of each type of strategy in the population.) Unfortunately, one also needs to assume that any two members of the population are equally likely to interact. While it may be true for sufficiently small populations of humans that any two interactions are equally likely, the plausibility of this decreases as the population size increases. By the time one reaches a population of the size of, say, New York, it certainly is no longer true that any two members are equally likely to interact.

Second, in his model of the Nash bargaining game, Skyrms only considers the effect of positive correlation among strategies. D'arms *et al.* point out that, while this makes sense for strategies which request at most half of the cake, this does not make sense for strategies asking for *more* than half of the cake. If one is going to introduce correlation into the model, one needs to allow for both positive and negative correlation. Negative correlation, in the Nash bargaining game, would correspond to some sort of avoidance behavior, where, say, individuals asking for six slices of the cake would try to steer clear of individuals asking for the same amount. D'Arms *et al.* construct a model, similar to Skyrms's, finding that when one allows for negative as well as positive correlation, the unfair polymorphisms reappear.

4. An agent-based, social network model

Description

In this section, I describe an agent-based social network model which improves upon Skyrms's replicator dynamic model, in the sense that it gives more robust results concerning the emergence of fair division for the Nash bargaining game. As before, for sake of simplicity, we assume that the cake is sliced into 10 pieces, and that individual requests are restricted to an integer number of slices. We replace the replicator dynamic assumption that we have an essentially infinite population by the assumption that the population P under consideration has only finitely many finitely many agents. Each agent in the population has a particular belief (or strategy) determining her behavior in the game of interest. Furthermore, we assume that an individual interacts only with those people who stand in some appropriate social relation to her. In general, these relations could be given by any connected graph whose nodes are the individual agents in the population. In this paper, I assume that the underlying social network has the form of a square lattice, where each agent is connected to some subset of the Moore 24 neighborhood. These relations are considered fixed since we assume individual beliefs change much faster than an agent's social relations. Finally, we assume that individuals follow some sort of imitative rule which determines how they change their beliefs over time.

The evolutionary dynamics used in this model are relatively common: at the start of each generation, each player receives a score equal to the number of slices of cake she receives when playing the appropriate game (either the Nash bargaining game or the ultimatum game) with her neighbors. At the end of each generation, an agent will change her strategy if some other agent in her neighborhood earned a higher score.

We allow agents to use one of four different update rules, each rule having a certain degree of plausibility.

The first update rule considered is "imitate the best neighbor." A very common update rule [see Nowak and May (1992; 1993), Lindgren and Nordahl (1994), Huberman and Glance (1993), and Epstein (1998)]. Each agent looks at her neighbors and mimics the strategy of the neighbor who did the best, where "best" means "earned the highest score." If ties occur, agents choose a random strategy by essentially flipping a coin.

The second update rule is "imitate with probability proportional to success." As before, each agent compares her score with those of her neighbors, modifying her strategy only if at least one neighbor did strictly better. However, instead of ignoring those neighbors who did better but not well enough to include their strategy in the set of highest-scoring ones, this update rule assigns to every neighbor who did better than an agent a nonzero probability that the agent will adopt her strategy. [For a formal description of this update rule, and the others, see Alexander (1999).]

The last two update rules are "imitate the strategy with the best expected payoff" and "adopt the best response strategy." Under the former rule, agents calculate the expected payoff of each strategy in their neighborhood, selecting the one with the highest value. With the latter rule, agents compute the best response strategy assuming that, in the next generation, none of their neighbors will change their strategies.

Obviously all of these update rules provide only rough approximations of the sort of rules real human agents would use. However, they are reasonable approximations in that they assume individual agents will use some rough-and-ready heuristic when determining what strategy to use in the next generation, instead of a computationally intensive optimization algorithm which only the most cognitively sophisticated could employ. Most of the time we find that the *simpler* update rules are the ones conducive to the evolution of norms. Perhaps our norms of fair division arose because they are the sorts of behaviors most beneficial to boundedly rational agents whose interactions are constrained by social networks.

The Nash bargaining game

Figure 2 illustrates the evolution of a 200×200 world in which all strategies are equally likely. Each block represents a single agent, and different shades of gray indicate how many slices of the cake each agent requests. Each slide portrays the state of the model after one generation, so the entire evolutionary process shown in figure 2 represents only nine generations. The dark gray color dominating the last few images corresponds to the strategy of fair division.

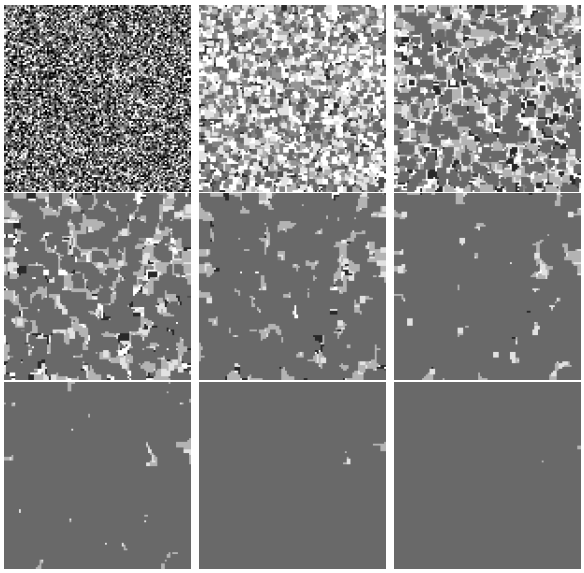

Figure 2: Fair division emerging from uniform random conditions

Figure 3 shows the ability of fair division to invade a world starting at one of the unfair Nash equilibria. A single agent adopting the strategy of fair division is sufficiently successful to initiate the spread of fair division throughout the population. Figure 4 illustrates the robustness of fair division in the presence of "mutation." In a world randomly initialized (with all strategies equally likely) with a mutation rate of 5%, fair division still becomes the dominant strategy with only the expected amount of mutational noise occurring in the background.

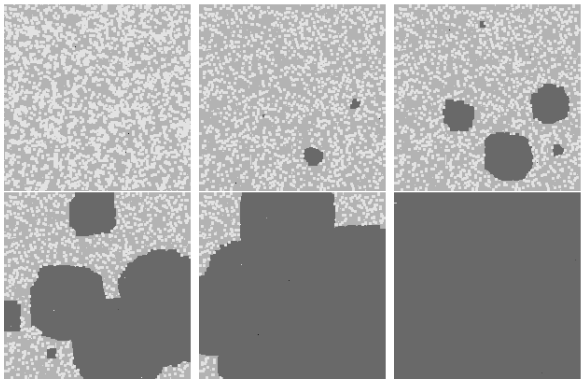

Figure 3: Fair division emerging from 4–6 polymorphism

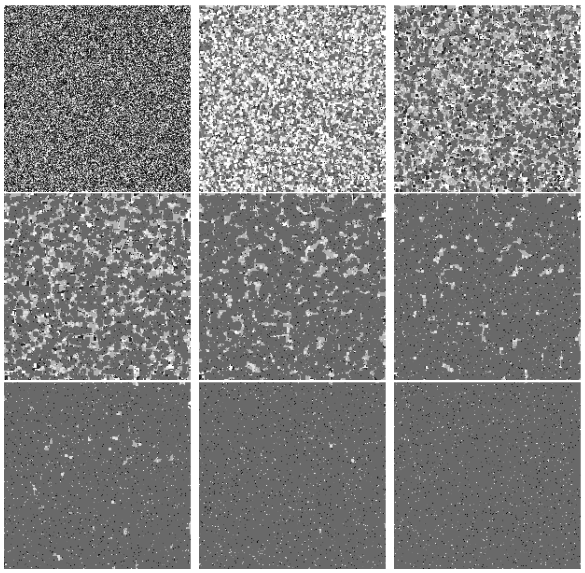

Figure 4: Fair division persisting in the face of mutations ($\mu = 0.05$)

Table 1 lists convergence results based on neighborhood and dynamic. In the column labeled "Dyn," "1" indicates the "imitate with probability proportional to success" update rule, "2" the "imitate best neighbor" update rule, and "3" the "imitate the strategy with best expected payoff." One can see that virtually no dependence on the neighborhood size or update rule exists, provided that we only consider agents using one of these three update rules. Figure 5 suggests, though, that using the "best response" update rule can give quite different results.

		Polymorphism						
Nbhd	Dyn	0–10	1–9	2–8	3–7	4–6	5	Other
VN	1	0	0	0	0	29	9970	1
	2	0	0	0	0	26	9966	8
	3	0	0	0	0	13	9984	3
M(8)	1	0	0	0	0	26	9973	1
	2	0	0	0	0	26	9908	66
	3	0	0	0	0	24	9970	6
M(24)	1	0	0	0	8	110	9879	3
	2	0	0	0	21	220	9721	38
	3	0	0	0	0	62	9934	4

Table 1: Convergence results based on neighborhood and dynamic

The reason for this odd behavior under the best response rule can be easily appreciated. Agents surrounded by a majority of neighbors who demand four slices of the cake will compute that the best-response strategy in the next generation is to request six slices of the cake. In a region consisting primarily of agents who request four slices, in the next generation all of those agents will request six slices. Of course, the best response strategy will then be to request four slices of the cake, and so on. Given the specification of the Nash bargaining game, this oscillatory behavior leads to a horribly *suboptimal* result for all of the agents, with a long-run average payoff of only two slices of cake; this is considerably less than what they would receive if they used a less sophisticated update rule (such as "imitate the best neighbor") and converged to a state where everyone asked for half of the cake.

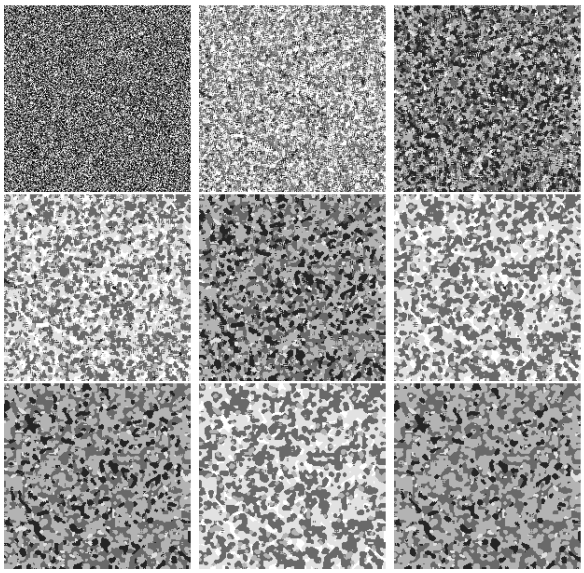

Figure 5: The disadvantageous best response update rule

The ultimatum game

The agent-based, social network model of the ultimatum game considered in this section is virtually the same as the model of the Nash bargaining game discussed in the previous section. The only relevant difference is that, on account of the slightly more complicated nature of the ultimatum game, we need to introduce an additional calculation in each generation. Before calculating each agent's score, for each interaction between two agents we randomly assign one agent the role of ultimatum giver and the other agent the role of ultimatum receiver. An agent's score equals the sum of the individual payoffs earned when that agent plays the ultimatum game with each of her neighbors, where each pairwise ultimatum game uses the assignment of roles made at the start of the generation.

Given the extent to which the agent-based, social network model increased the probability of fair division emerging for the Nash bargaining game, one might ex-

pect a similar effect to occur for the ultimatum game. (In the ultimatum game, the analogous effect would be to have the evolutionary dynamics cause the Fairman and/or Easy Rider strategy to acheive predominance.) However, it turns out that in the agent-based model described above the Fairman strategy *fails* to dominate in the vast majority of cases.[8]

Figure 6 illustrates how, under the "imitate best neighbor" update rule, the Gamesman and Mad Dog strategies dominate. Since we are now considering the underlying game to be the ultimatum game, rather than the Nash bargaining game, it should be noted that the color scheme used in figures 6, 7, and 8 has a different meaning than before. In these three figures, black represents the Gamesman strategy, and the darker gray color, Mad Dog. In the final image of figure 6, the only surviving strategies are Gamesman and Mad Dog.

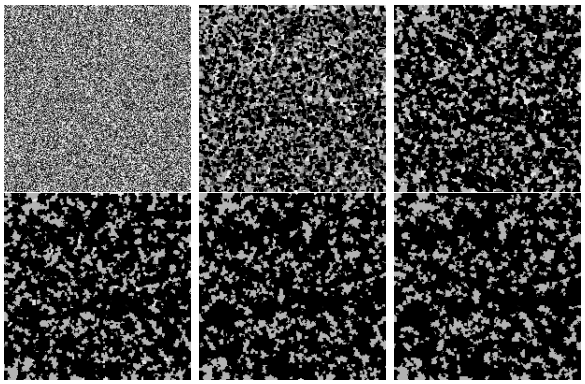

Figure 6: The emergence of gamesman and mad-dogs

Under the update rule of "imitate the best neighbor," the Fairman strategy lacks the robustness properties possessed by the demand-half strategy in the Nash bargaining game. For example, figure 7 shows how a pure population of Fairman (indicated by light gray) can eventually be invaded and overwhelmed by Gamesmen if we allow a very small amount of mutation. Since the sequence of figure 7 was not sampled at constant time intervals (unlike the rest of the figures in this paper), I indicate the exact generation of each image explicitly.

[8]In the context of the ultimatum game, by a "fair-playing" strategy I mean either Fairman or Easy Rider, since these are the only two self-consistent strategies which offer half of the cake. I take the other two strategies which offer half of the cake to be ones which we would not expect rational agents to adopt, since they are not self-consistent—they refuse the very offers they make! (See the table in section .)

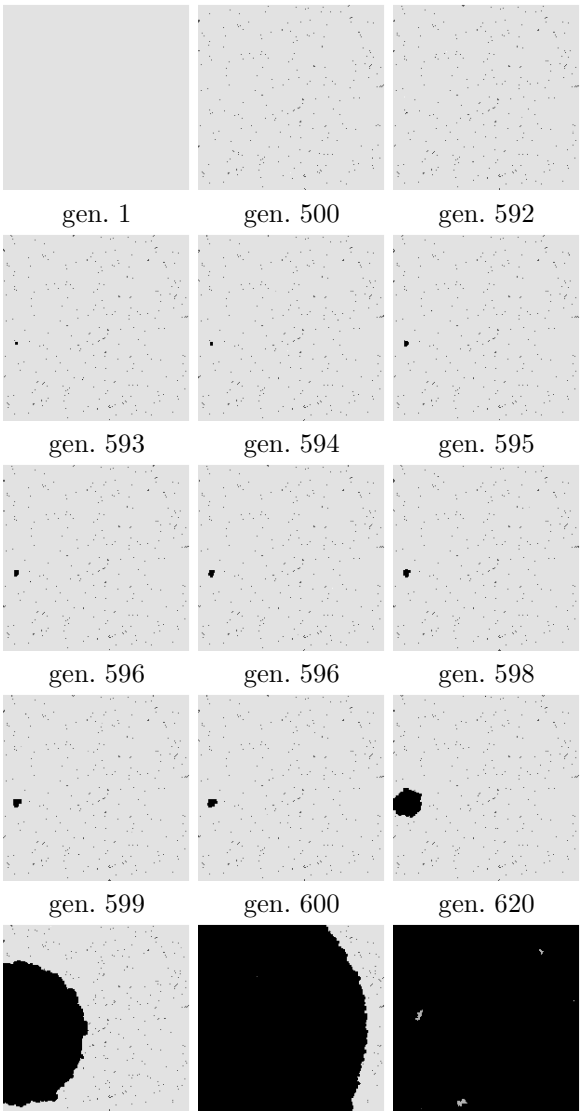

Figure 7: The death of Fairman

During the first 500 generations, the only mutant strategies which survive in a pure Fairman population are Easy Riders. By the 500th generation, though, enough Easy Riders mutants have appeared to allow a Gamesman mutant to flourish. Once a reasonably sized Gamesman cluster has appeared (which occurs in the sequence of figure 7 by generation 593), Gamesmen may spread into regions occupied by Fairman without needing to piggyback on the presence of Easy Riders.[9] Over time, the population will eventually arrive at a state consisting primarily of Gamesmen, with a few Mad Dogs.[10]

However, if we modify the game slightly, allowing Fairmen to "punish" agents who do not make fair offers, we find that Fairmen, who previously became extinct in a few generations, may persist in the hostile environment created by the presence of Gamesmen. Figure 8 illustrates the evolution of a population in which the Fairman strategy punishes greedy neighbors (which, in this context, means merely that when a "greedy" strategy interacts with the Fairman strategy, they receive a negative payoff). The presence of the Fairman strategy, represented by light gray, shows how the strategy corresponding to our norm of fairness persists when the parameter controlling the severity of the punishment exceeds a certain value.

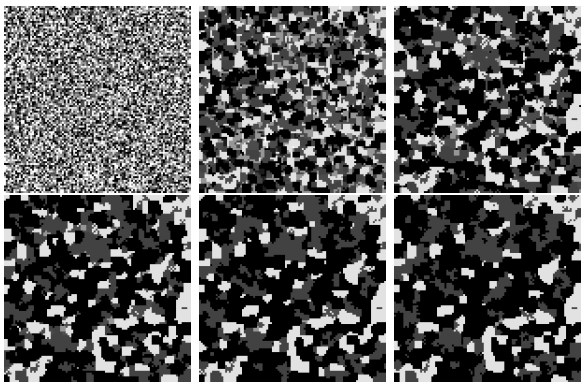

Figure 8: The emergence of fair play in the ultimatum game

[9]Since Easy Riders accept any offer, a Gamesman mutant appearing in a region with a high concentration of Easy Riders may exploit their presence to obtain a critical foothold in the world. The random assignment of roles may fall in favor of the Gamesman mutant, meaning that he will make offers to Easy Riders and receive offers from Fairman. When this happens, the Gamesman mutant will typically earn a score high enough to persist into the next generation, spreading his strategy to neighboring agents.

[10]Mad Dogs appear by mutation and may survive in a pure Gamesman population since, when no other strategies are present, they are indistinguishable from Gamesmen.

6. Conclusion

There exists a rich philosophical tradition which seeks to ground norms of moral and political obligation in the self-interested actions of individual agents. The models of this paper demonstrate that the emergence of norms can depend on dynamical considerations which are not immediately apparent. In particular, these models suggest that our norm of fair division in games having the structure of the Nash bargaining game depends on the constraint of some underlying social network on our interactions. In addition to this constraint, in the ultimatum game the models suggest that acquiring the ability to punish "deviant" strategies plays an essential role in the development of our concept of justice. The primary point is that evolutionary accounts of norms which neglect to take seriously the underlying dynamics, and the structure of the underlying game, do so upon risk of descriptive inaccuracy and hence do not provide a plausible account of normativity.

References

Jason M. Alexander. The (spatial) evolution of the equal split. Technical report, Institute for Mathematical Behavioral Sciences, U.C. Irvine, 1999.

Martin Barrett, Ellery Eells, Branden Fitelson, and Elliott Sober. Models and reality—a review of Brian Skyrms' *Evolution of the Social Contract*. *Philosophy and Phenomenological Research*, 59(1):237–241, March 1999.

K. Binmore, A. Shaked, and J. Sutton. Testing noncooperative bargaining theory: A preliminary study. *The American Economic Review*, 75(5):1178–1180, December 1985.

Justin D'Arms, Robert Batterman, and Krzyzstof Górny. Game theoretic explanations and the evolution of justice. *Philosophy of Science*, 65:76–102, March 1998.

Joshua A. Epstein. Zones of cooperation in demographic prisoner's dilemma. *Complexity*, 4(2):36–48, 1998.

W. Güth and R. Tietz. Strategic power versus distributive justice. An experimental analysis of ultimatum bargaining. In H. Brandstätter and E. Kirchler, editors, *Economic Psychology. Proceedings of the 10th IAREP Annual Colloquium*, pages 129–137. Rudolf Trauner Verlag, Linz, 1985.

Werner Güth and Reinhard Tietz. Ultimatum bargaining behavior: A survey and comparision of experimental results. *Journal of Economic Psychology*, 11:417–449, 1990.

Werner Güth, Rolf Schmittberger, and Bernd Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3:367–388, 1982.

Bernardo A. Huberman and Natalie S. Glance. Evolutionary games and computer simulations. *Proc. Natl. Acad. Sci.*, 90:7716–7718, August 1993.

Philip Kitcher. Games social animals play: Commentary on Brian Skyrms' *Evolution of the Social Contract*. *Philosophy and Phenomenological Research*, 59(1):221–228, March 1999.

David M. Kreps. *Game Theory and Economic Modelling*. Oxford University Press, 1990.

Kristian Lindgren and Mats G. Nordahl. Evolutionary dynamics of spatial games. *Physica D*, 75:292–309, 1994.

R. Duncan Luce and Howard Raiffa. *Games and Decisions: Introduction and Critical Survey*. John Wiley and Sons, Inc., 1957.

John F. Nash. The bargaining problem. *Econometrica*, 18:155–162, 1950.

Janet Neelin, Jugo Sonnenschien, and Matthew Spiegel. A further test of noncooperative bargaining theory: Comment. *The American Economic Review*, 78(4):824–836, September 1988.

Martin A. Nowak and Robert M. May. Evolutionary games and spatial chaos. *Nature*, 359:826–829, October 1992.

Martin A. Nowak and Robert M. May. The spatial dilemmas of evolution. *International Journal of Bifurcation and Chaos*, 3(1):35–78, 1993.

R. V. Nydegger and G. Owen. Two-person bargaining: An experimental test of the Nash axioms. *International Journal of Game Theory*, 3(4):239–249, 1974.

Jack Ochs and Alvin E. Roth. An experimental study of sequential bargaining. *The American Economic Review*, 79(3):355–384, June 1989.

Alvin E. Roth, Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Toyko: An experimental study. *The American Economic Review*, 81(5):1068–1095, December 1991.

Alvin E. Roth. Bargaining experiments. In J. Kagel and A. Roth, editors, *The Handbook of Experimental Economics*, chapter 4, pages 253–348. Princeton University Press, 1995.

A. Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50:1123–42, September 1982.

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

Peter D. Taylor and Leo B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40:145–156, 1978.

Richard H. Thaler. Anomalies: The ultimatum game. *Journal of Economic Perspectives*, 2(4):195–206, 1988.

Menachem E. Yaari and Maya Bar-Hillel. On dividing justly. *Social Choice and Welfare*, 1:1–24, 1984.